Molecular insights into the mating system of the marine diatom Pseudo-nitzschia multistriata using genetic and genomic approaches

Thesis

Molecular insights into the mating system of the marine diatom *Pseudo-nitzschia multistriata* using genetic and genomic approaches

Laura Vitale

Thesis submitted for the degree of

Doctor of Philosophy (PhD)

in Life and Biomolecular Sciences

December 2015

Open University, London

Stazione Zoologica Anton Dohrn, Napoli

The Open University

DATE OF SUBMISSION: 30 DECEMBER 2015
DATE OF AWARD: 4 NOVEMBER 2016

CONTAINS PULLOUTS

Director of Studies: Dr Marina Montresor

Stazione Zoologica Anton Dohrn, Napoli, Italy

Internal Supervisor: Dr Maria Immacolata Ferrante

Stazione Zoologica Anton Dohrn, Napoli, Italy

External Supervisor: Prof Wim Vyverman

Ghent University, Ghent, Belgium

Abstract

Sexual reproduction is a fundamental phase in the life cycle of diatoms, linked to the production of genotypic diversity and the formation of large-sized initial cells that ensure population persistence. It occurs only within cells below a certain threshold size and, in heterothallic diatoms, only between strains of opposite mating types.

We aim at identifying genes involved in mating type determination in the marine planktonic diatom *Pseudo-nitzschia multistriata*. This species is recorded in coastal waters worldwide and produces the neurotoxin domoic acid. A reference genome has been generated and transcriptomes have been produced for strains of opposite mating type (MT+ and MT-).

Differential expression analysis provided a list of candidate MT-biased genes validated with qPCR. Four MT-biased genes were identified, two in MT+ and two in MT-. The expression pattern of the candidate genes was followed in a 24 hours' time course experiment to verify whether they were regulated in dependence of light or cell cycle phases. Experimental evidences demonstrated their involvement during mating recognition in early stages of sexual reproduction while preliminary genetic analyses excluded that they could be the master gene responsible for mating type determination. The description of the four genes was improved through computational characterization to understand their role in the chemical communication occurring between opposite mating types. A further step towards the identification of the MT locus will include a Bulked Segregant Analysis applied to a library of 30 MT+ and 30 MT- F1 strains obtained through DNA deep sequencing.

Elucidating the molecular and genetic basis of MT determination and sexual reproduction in diatoms will contribute to a better understanding of the regulation and evolution of their life cycles and reproductive strategies. Results from this study could also provide

Director of Studies: Dr Marina Montresor

Stazione Zoologica Anton Dohrn, Napoli, Italy

Internal Supervisor: Dr Maria Immacolata Ferrante

Stazione Zoologica Anton Dohrn, Napoli, Italy

External Supervisor: Prof Wim Vyverman

Ghent University, Ghent, Belgium

Acknowledgments

Firstly, my thanks go to the Stazione Zoologica Anton Dohrn of Naples for funding my PhD project. I would like to express my gratitude to my Director of study Marina Montresor for the continuous support at my PhD research, for her patience and motivation. Her guidance helped me to grow in science. I would like to thank my internal supervisor Mariella Ferrante who provided me the opportunity to join her team. Without her precious support it would not be possible to conduct this research and write this thesis. My thanks also go to my external supervisor Wim Vyverman for his helpful suggestions encouragement and positivity. A special thank goes to my third party monitor Christophe Brunet. I would especially like to thank the bioinformatic group of Remo Sanges, with Swaraj Basu and Francesco Musacchia for the all the analyses conducted and for the friendly support. I would like to thank the Molecular Biology and Bioinformatics Unit for sharing with me their technical skills. I also want to thank all the persons that supported me in different ways: in particular, Monia Russo and Valeria Sabatino for molecular support, Lazaro Marin Guirao, Greta Busseni and Luigi Caputi for statistical support, Carmen Minucci for technical support in BSA experiment, Eleonora Scalco and Laura Escalera for teaching me how to tame, domesticate and breed *Pseudo-nitzschia multistriata* and for the bibliographical help, Maria Paola Tomasino for helping me in creating heatmap, Roberta Piredda for her wise advices on mapping procedures and Arianna Smerilli for the stylish support. My thanks also go to the Genome Analysis Centre (TGAC) in Norwich (UK) for sequencing *Pseudo-nitzschia multistriata* genome and to the Joint Genome Institute (JGI) for funding *Pseudo-nitzschia multistriata* transcriptome sequencing.

Finally, I wish to thank all the friends I met in the lab for all the fun we had in the last three years being key part of my life; all my friends far from the world of research, Rosanna, Laura, Miele's ones, Fulvia, and Angela which helped me to keep contact with reality. Last

but not least my family, Mamma, Papà, Emy, Semola and Cenere, I owe to them what I am and without whom I could not have been able to achieve this goal.

Table of contents

List of tables and figures

Chapter 1

General introduction

1.1 Diatoms

Diatoms are a key group of unicellular eukaryotes belonging to the super-group Chromalveolata, first-rank group Stramenopiles, second-rank group Bacillariophyta following the rank organization proposed by Adl *et al.* (2005). The etymology of the word 'diatom' derives from the Greek 'dia tomos' meaning 'cut in half' and reflects the particular structure of their mineral covering, the frustule that is constituted by two, slightly unequal, parts that fit together as that of a lid on a box. These two parts are called 'epi-theca' and 'hypo-theca', and each of them is constituted by a valve and a series of cingular bands. The frustule surfaces are perforated with minute pores that allow dissolved molecules to enter into the cell. The frustule is made essentially of hydrated silicon ($SiO_2.nH_2O$). The silicon dissolved in sea water as silicic acid, enters the diatom cell thanks to transporter proteins. The polymerization of silica takes place within specialized intracellular compartments, the silica deposition vesicles (SDV). Through this mechanism of biomineralization, diatoms control the biogenic cycling of silicon in the world's oceans (Treguer *et al.*, 1995).

Astonishing diversity characterizes this group that ranges in size from a few micrometres to a few millimetres and exists either as single cells or as chains of connected cells (Kooistra *et al.*, 2007). The approximate number of morphologically-defined marine diatom species is about 1,800 (Sournia *et al.*, 1991), but evidences provided by metabarcode studies (de Vargas *et al.*, 2015) and inferences based on the increasing evidence of cryptic and pseudo-cryptic species (Mann & Vanormelingen, 2013) show that these figures are largely underestimated. Diatoms are important members of both planktonic and benthic phytoplankton communities and play a fundamental role in the biogeochemical cycles of the global oceans generating most of the organic matter that serves as food for life in the sea. They carry out one-fifth of the global carbon fixation that is as much organic carbon as all the terrestrial rainforests combined (Armbrust, 2009).

The threshold for diatom growth is set by the availability of light and nutrients. The most prominent recurrent feature of the seasonal plankton cycle in temperate and boreal systems is the spring bloom that takes place when day-length increases and remineralized inorganic nutrients become available in the upper portion of the water column after the deep winter mixing. The spring bloom is generally dominated by comparatively few species of unrelated genera of diatoms. Typical examples of recurrent bloomers are species of the cosmopolitan diatom genera *Skeletonema*, *Thalassiosira*, and *Chaetoceros* (Assmy & Smetacek, 2009). When the blooms are over (generally due to the exhaustion of nutrients or by natural death of the populations), diatoms sink along the water column and export the organic carbon in the deep layers of the ocean, where it provides the food for benthic organisms and is respired by the bacteria (Smetacek, 1999). Diatoms thus greatly influence global climate, atmospheric carbon dioxide concentration and marine ecosystem function, and understanding the biology of these important organisms and how they will respond to the rapidly changing conditions of the oceans is critical to predict the future health of the environment.

Diatoms originated more than 250 Myr ago by a secondary endosymbiotic event in which a heterotrophic eukaryote engulfed a red alga (Armbrust, 2009). Over time the red alga transformed into plastids of the heterokont and gene transfer continued from red algal genome and plastid to the host genome (Armbrust *et al.*, 2004). Red algae themselves originated by a primary endosymbiosis event occurred about 1.2 billion years ago when a heterotrophic eukaryote engulfed a photosynthetic cyanobacterium to give rise to an ancestor of the 'group plantae' (Yoon *et al.*, 2004). Diatoms genomes support the idea of secondary endosymbiosis, but gene analysis scored a complex combination of genes and pathways acquired from a variety of sources. In *Phaeodactylum tricornutum* 170 genes of clear red algal origin were found but many more were of green algal and bacterial origin (Bowler *et al.*, 2008). The endosymbiotic events play a defined role in the overall

capabilities of diatoms, but subsequent gains (or losses) of specific genes, largely from bacteria, presumably helped to adapt to new ecological niches (Armbrust, 2009).

Diatoms are divided in four groups, which differ for morphological and phylogenetic traits. Molecular phylogenetic studies conducted on the small subunit of the ribosomal rDNA (SSU-rDNA) highlighted the evolutionary sequence of the four groups. The radial centric diatoms are the most primitive lineage, followed by bi- and multipolar centric, than araphid pennates and finally raphid pennates, which are the youngest lineage (Kooistra *et al.*, 2007). Marine planktonic diatoms mostly belong to the centric lineages, but there are some genera of pennate diatoms, e.g. *Pseudo-nitzschia, Fragilariopsis, Thalassiothrix, Asterionellopsis* that have a planktonic habit and are an important component of the marine plankton. A metabarcode study on samples collected within the Tara-Oceans expedition, indicated once more that *Chaetoceros* and *Thalassiosira* are the most abundant genera, followed by *Corethron, Fragilariopsis, Actinocyclus, Leptocylindrus,* and *Pseudo-nitzschia* (Malviya *et al.*, 2016). If we compare these data with the list of species for which information on the occurrence of sex is available (Table 1.1), we realize that knowledge is notably patchy and that we completely miss information for some of the most important genera. Table 1.1 includes data from publications in which a solid documentation was provided on sexual stages, both in case of observations with culture material in the laboratory and for studies in the natural environment.

Table 1.1: List of marine planktonic diatom species for which information on the occurrence of a sexual phase is available. 'Culture' indicates that evidence for sex was provided by observations and experiments carried out in the laboratory with culture material; 'Nature' indicates that information on sexual reproduction was provided by observations of sexual stages in natural populations at sea.

Species	References	Evidence for sex
Actinocyclus ehrenbergii	von Stosch and Drebes (1964)	Culture
Actynoptychus senarius (as *A. undulatus*)	von Stosch and Drebes (1964)	Culture
Aulacodiscus argus	von Stosch and Drebes (1964)	Culture
Attheya decora	Drebes (1977a)	Culture
Bacteriastrum hyalinum	Drebes (1972)	Culture
Chaetoceros curvisetus	Furnas (1985)	Culture

C. diadema	Hargraves (1972); French III and Hargraves (1985)	Culture
C. dichaeta	Assmy *et al.* (2008)	Culture
C. didymus	von Stosch *et al.* (1973); Drebes (1977b)	Culture
C. eibenii	von Stosch *et al.* (1973)	Culture
C. laciniosus	Jensen *et al.* (2003)	Culture
C. protuberans	(Lechuga-Devéze & Hernández-Becerril, 1988)	Culture
Cylindrotheca closterium	Vanormelingen *et al.* (2013)	Culture
Corethron pennnatum (as *C. criophilum*)	Crawford (1995)	Nature
Coscinodiscus concinnus	Holmes (1966); Drebes (1977b)	Culture
Coscinodiscus granii	Schmid (1995)	Culture
Coscinodiscus wailesii	Nagai and Manabe (1994); Nagai *et al.* (1996); Jensen *et al.* (2003)	Culture
Ditylum brighwellii	Koester *et al.* (2007)	Culture
Eucampia zodiacus	Nishikawa *et al.* (2013)	Nature
Fragilariopsis kerguelensis	Assmy *et al.* (2006) Fuchs *et al.* (2013)	Nature Culture
Haslea karadagensis	Davidovich *et al.* (2012)	
Haslea ostrearia	Davidovich *et al.* (2009); Mouget *et al.* (2009); Gastineau *et al.* (2013)	Culture
Helicotheca tamesis (as *Streptotheca tamesis*)	von Stosch (1954)	Culture
Leptocylindrus danicus	French III and Hargraves (1985); Nanjappa *et al.* (2013)	Culture
Leptocylindrus hargravesii	Nanjappa *et al.* (2013)	Culture
Lithodesmiulm undulatum	Manton and von Stoch (1966); Manton *et al.* (1969)	Culture
Melosira moniliformis	Migita (1967b); von Stosch (1958)	Culture
Melosira moniliformis v. *octagona*	Idei and Chihara (1992)	Culture
Neodenticula seminae	Kurihara and Takahashi (2002)	Nature
Nitzschia longissimi	Kaczmarska *et al.* (2007)	Culture
Odontella granulata (as *Biddulphia granulata*)	von Stosch (1956); Hoppenrath *et al.* (2009)	Culture
Odontella mobiliensis (as *Biddulphia mobiliensis*)	von Stosch (1954)	Culture
Odontella regia	Hoppenrath *et al.* (2009); Hegde *et al.* (2011)	Culture/Nature
Odontella rhombus (as *Biddulphia rhombus*)	von Stosch (1956)	Culture
Odontella sinensis	von Stosch (1956); Hoppenrath *et al.* (2009)	Culture/Nature

Pseudo-nitzschia arenysensis	Quijano-Scheggia *et al.* (2009b)	
Pseudo-nitzschia arenysensis (as *P. delicatissima*1)	Amato *et al.* (2005); Amato *et al.* (2007); Levialdi Ghiron *et al.* (2008)	Culture
P. australis	Holtermann *et al.* (2010)	Nature
P. brasiliana	Quijano-Scheggia *et al.* (2009a)	Culture
P. calliantha	Amato *et al.* (2007)	Culture
P. cf. calliantha	Sarno *et al.* (2010)	Nature
P. cuspidata	Amato *et al.* (2007); Lundholm *et al.* (2012)	Culture
P. delicatissima	Kaczmarska *et al.* (2008)	Culture
P. cf. delicatissima	Sarno *et al.* (2010)	Nature
P. dolorosa	Amato *et al.* (2007)	Culture
P. fraudulenta	Chepurnov *et al.* (2004)	Culture
P. mannii	Amato and Montresor (2008)	Culture
P. multiseries	Davidovich and Bates (1998); Hiltz *et al.* (2000); Kaczmarska *et al.* (2000)	Culture
P. multistriata	D'Alelio *et al.* (2009); D'Alelio *et al.* (2010); (Scalco *et al.*, 2014)	Culture/Nature
P. pseudodelicatissima	Davidovich and Bates (1998); Amato *et al.* (2007)	Culture
P. pungens	Chepurnov *et al.* (2005); Casteleyn *et al.* (2009); Holtermann *et al.* (2010)	Culture/Nature
P. pungens var. *aveirensis*	Churro *et al.* (2009)	Culture
P. subcurvata	Fryxell *et al.* (1991)	Culture
Skeletonema costatum	Migita (1967a); Davis *et al.* (1973); Gallagher (1983)	Culture
S. marinoi	Godhe *et al.* (2014)	Culture
Stephanopyxis palmeriana	Drebes (1966)	Culture
Stephanopyxis turris	Drebes (1964); von Stosch (1956)	Culture
Thalassiosira angulata	Mills and Kaczmarska (2006)	Culture
T. eccentrica	Drebes (1979)	Culture
T. punctigera	Chepurnov *et al.* (2006)	Culture
T. weissflogii	Vaulot and Chisholm (1987); Armbrust *et al.* (1990); Armbrust (1999); von Dassow *et al.* (2006)	Culture

1.2 Diatom life cycle

Diatoms have a peculiar life cycle characterized by a dominant diploid (2N) vegetative phase in which they undergo mitotic division and a short sexual phase, including meiosis, gametogenesis and fertilization, during which gametes are the only haploid (N) stage produced. This life cycle is called diplontic (Round *et al.*, 1990).

The life cycle of diatoms is characterized by a strong link with the cell size. Cells can undergo sexual reproduction only in a defined size window, the sexualisation size threshold (SST). Cell size window for sexual reproduction can vary amongst different species. For the majority of diatoms, sexuality is an obligatory phase of their life cycles, but there are some species that restore their size trough vegetative cell enlargement or uniparental auxosporulation (Kaczmarska *et al.*, 2013). The modality through which sexual reproduction occurs has been studied in the laboratory for several diatoms, mostly freshwater benthic species (Chepurnov *et al.*, 2004). Generally in diatoms, the restoration of the large cell size in a given population is accomplished through sexual reproduction, as the zygote is the only stage that does not have a silica wall and can therefore expand to the maximum cell size. A few exceptions apart, sexual reproduction is an obligate phase in diatom life cycles, required not only to increase genetic diversity but also to escape the miniaturisation process that would eventually lead to extinction of the population.

The life cycle of many diatoms also includes the formation of resting stages, either spores, which are morphologically differentiated from vegetative cells and have a thick wall, or resting cells, which are morphologically similar to vegetative cells, but physiologically differentiated (Hargraves, 1976). Resting stages can be quiescent in the sediments for years and might serve as survival stages for the population during adverse conditions for growth (McQuoid & Hobson, 1996).

Before undergoing vegetative division, diatoms must increase about twofold their cell volume, double the mitochondria, plastids and other organelles, replicate the chromosomes

and then segregate full components in each daughter cell. This increase in cytoplasm material is associated with an increase in cell size (Round *et al.*, 1990). Mitotic division creates two daughter cells that inherit one half of the frustule from the parental cell – that becomes the external theca - and synthesize *ex novo* the internal thecae. It follows that one daughter cell has the same size as the parental cell and the other one is smaller; this cell division modality causes a gradual decrease in the average cell size - length in pennates and diameter in centrics - of the population (Round, 1972). The large majority of diatoms escape this progressive miniaturization through sexual reproduction during which large-sized cells are formed within a specialized and flexible zygote, the auxospore. The zygote lacks the rigid siliceous wall, so it is free to expand and form the auxospore. The auxospore starts expanding by the formation of a composite organic-siliceous wall, called perizonium, consisting of bands made of an organic matrix in which some silica is incorporated. After the auxospore has reached the maximum species-specific size, the two valves of the initial cell are formed.

In the two main groups of diatoms, centric and pennate, two very different modalities of sexual reproduction are present (Fig1.1) (Mann *et al.*, 1999, Chepurnov *et al.*, 2004).

Centric diatoms generally have a homothallic mating system, i.e. gametes of opposite mating type (+) and (-) are produced in the same clonal culture, and they have an oogamous sexual reproduction. Within the '-' (female) gametangium they form one or two sessile macro-gametes (egg cell/s) and within the '+' (male) gametangium, they produce numerous small uni-flagellate gametes (sperm cells). In most species, only one haploid egg is produced while spermatogenesis starts with a series of special mitotic divisions, during which cells do not expand, that brings to a progressive reduction in cell size and plastid number per cell (Drebes, 1977b, Chepurnov *et al.*, 2004). After the spermatozoids are released, they swim actively towards the egg cell, but the mechanisms of attraction and recognition between sperms and eggs in centric diatoms are unknown.

Pennate diatoms generally have a heterothallic mating system, i.e. sexual reproduction occurs only when mixing strains of opposite mating type, (-) and (+). Most raphid diatoms are isogamus, i.e. gametes are similar in shape and size but functionally distinct (Round *et al.*, 1990, Chepurnov *et al.*, 2004). The gametes of pennate diatoms are non-flagellate and have limited capacity of movement, so the two gametangia must be positioned close enough to allow conjugation. In fact, interaction between opposite mating types is required to start meiosis and gametogenesis. In pennate diatoms, only one or two gametes are produced for each gametangium and the number of gametes is the same for both mating types (Round *et al.*, 1990, Chepurnov *et al.*, 2004).

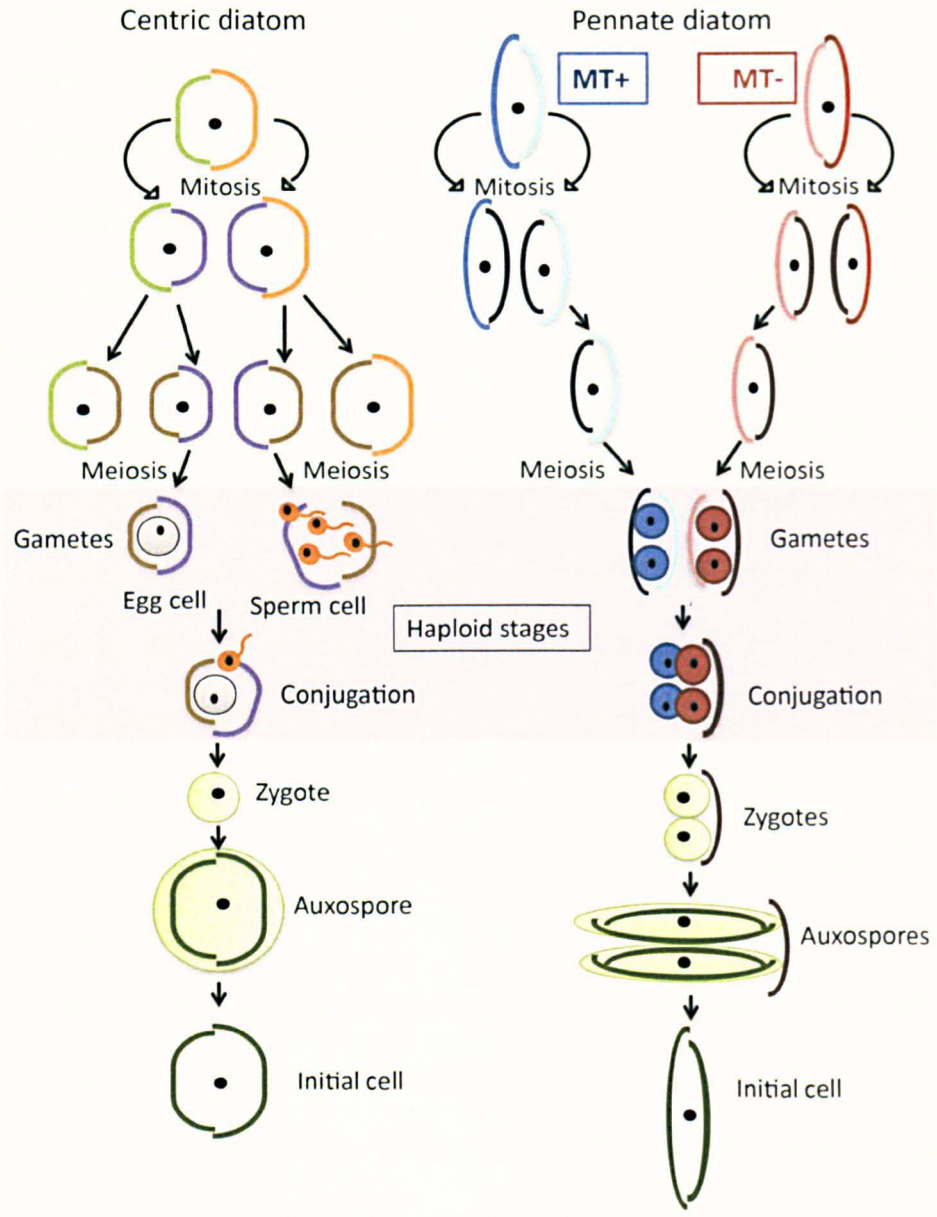


Figure 1.1: Schematic drawing of the life cycle of a centric and a pennate diatom. Diatom cells are diploid and are surrounded by a rigid frustule made of two unequal thecae. During mitosis, the new thecae are synthesized inside the maternal frustule. This causes a progressive decrease in the population cell size. The formation of gametes takes place following meiosis in cells (gametangia) that are below a species-specific size threshold for sexualisation. In centric diatoms, large macrogametes (egg cells) and small uniflagellated microgametes (sperm cells) are produced within the same strain. In pennate diatoms, the formation of gametes occurs when two strains of opposite mating type are in close contact; gametangial cells pair side to side and meiosis takes place. Conjugation of the haploid gametes produces a zygote that expands into an auxospore. Within the auxospore, the large initial cell is synthesized (Montresor *et al.*, 2016).

Information on the molecular mechanisms that regulate transitions among different life cycle phases of unicellular microalgae (von Dassow & Montresor, 2011) and aspects of the

sexual phase in protists, i.e. mating systems, pheromone signalling and gamete conjugation, is limited to studies carried out on a handful of model species (e.g., (Umen, 2011); (Sekimoto *et al.*, 2012); (Goodenough & Heitman, 2014)) and are almost unknown. Nevertheless, the increased availability of molecular techniques and genomic resources starts providing insights into the life cycle of protists (von Dassow *et al.*, 2009, Grimsley *et al.*, 2010) allowing now to infer the presence of sex also in microalgae for which experimental evidence is still lacking. This is the case of the small prasinophyte *Ostreococcus tauri,* where evidence of recombination and chromosomal segregation was detected analysing eight loci on neutrally evolving intergenic regions (Grimsley *et al.*, 2010), or the symbiontic dinoflagellate *Symbiodinium,* in which meiosis-specific genes have been detected (Chi *et al.*, 2014).

In evolutionarily older centric diatoms, gamete differentiation is generally induced by environmental cues, once the appropriate size for sex has been reached. The range of factors potentially involved in triggering this process is broad, providing a complex picture from which it is difficult to extract common rules. In heterothallic pennate diatoms the species-specific cell size window is the primary factor that allows sexualisation however the sexual phase appears to be regulated by endogenous factors and evidence is building up for the role of sex pheromones in governing the perception of opposite mating types and inducing gametogenesis (Frenkel *et al.*, 2014). The first experimental evidence for the presence of sex pheromones has been gained for the freshwater diatom *Pseudostaurosira trainorii,* where female cells secrete a pheromone that induces the sexualisation of male cells. These male cells, in turn, secrete a pheromone that induces the sexualisation of female cells, which attract the male gametes (Sato *et al.*, 2011). A sexual phase mediated by sex pheromones has been described also for the benthic diatom *Seminavis robusta* (Gillard *et al.*, 2013, Moeys *et al.*, 2016). The pheromone system costs of a sex-inducing pheromone (SIP+), secreted by MT+, that triggers the switch from mitosis-to-meiosis in

MT− and induce the transcription of proline biosynthesis genes. The female sexualized cells produce the pheromone L-diproline that attracts male cells and probably induces the expression of a diproline receptor on their surface, thus allowing the formation of gametangial pairs (Moeys *et al.*, 2016). The production of sex pheromones has not been proven for marine planktonic diatoms yet. However, there is evidence for a density-dependent mechanism of sexualisation in *Pseudo-nitzschia multistriata*, where a cell concentration threshold is required for sex to occur (Scalco *et al.*, 2014). This may be evidence for a density-dependent mechanism that controls the production/perception of chemical signals. This hypothesis is corroborated also by the arrest of vegetative growth of the population in concomitance with sexualisation (Scalco *et al.*, 2014), a mechanism reported also for yeast (e.g., (Merlini *et al.*, 2013)) that should have the function of synchronizing the population in the G1 cell cycle phase in which cells are receptive to sex pheromones. A positive correlation between cell concentration and number of auxospores has been detected also in *Skeletonama marinoi* (Godhe *et al.*, 2014). Chemical compounds structurally similar to the sex-pheromone ectocarpene of brown algae have been characterized from freshwater diatoms (Derenbach & Pesando, 1986, Pohnert & Boland, 2002), thus suggesting that these compounds might have a similar role in diatoms that belong to the same Stramenopile clade.

1.3 Sex determination

1.3.1 Sex determination systems

The awareness that many of the mechanisms critical to basic animal development have been conserved across more than 500 million years of evolution is revolutionary. But not all developmental processes are conserved; an outstanding example is sex determination (Haag & Doty, 2005). A sexual population generally consists of two sex which are determined genetically by a pair of sex chromosomes or by environmental cues (Bergero & Charlesworth, 2009). Indeed, the two broadest categories of sex determination are:

Genetic sex determination (GSD), in which the sex of offspring is set by a sex chromosome or a MT-locus (mating type locus). Sexual identity is governed by sex chromosomes in plants and animals, and by mating type (MT) loci in fungi and unicellular eukaryotes.

Environmental sex determination (ESD), in which sex is determined by temperature (as in turtles), local sex ratio (as in some tropical fish), or population density (as in mermithid nematodes).

Sexual differentiation is common in eukaryotic organisms from yeasts to humans. In the following, I will present examples of genetic sex determination (GSD) systems.

Vertebrates

In the animal kingdom, there are different mechanisms for sex determination, with the predominant ones based on the presence of distinct chromosomes leading to oocyte-producing females and sperm-producing males (Zanetti & Puoti, 2013). The most common type of sex determination in vertebrates involves sex chromosomes. If the male is the sex with two different sex chromosomes (male heterogamety), the sex chromosomes are referred to as X and Y: females are XX, males are XY. Likewise, if the female is the sex

with two different sex chromosomes (female heterogamety), the sex chromosomes are Z and W: females are ZW, males are ZZ. Sex determination by sex chromosomes is universal in birds (female heterogamety ZW) and mammals (male heterogamety XY) and is present in both forms (male and female heterogamety) among reptiles, amphibians, and fishes.

Invertebrates: *Drosophila* and *Caenorhabditis*

In the model fly *Drosophila*, sex determination is achieved by a balance of female determinants on the X chromosome and male determinants on the autosomes. Normally, flies have either one or two X chromosomes and two sets of autosomes. If there is only one X chromosome in a diploid cell (1X:2A), the fly is male. If there are two X chromosomes in a diploid cell (2X:2A), the fly is female. A quantitative chromosomal signal, the X:A ratio, decides whether the key gene in sex determination, *SXl* (Sex lethal 1) is active (XX) or inactive (XY). The functional state, ON or OFF, of *SXl*, regulated via a few subordinate regulatory genes, controls a switch gene (*DSX*) that can express two mutually exclusive functions, M (male) or F (female). These serve to repress either the female or the male set of differentiation genes, thus directing the cells either into the male or into the female sexual pathway (Baker & Belote, 1983).

As in *Drosophila,* also the nematode *Caenorhabditis elegans* has an XX/XO sex chromosome system, but in its genome the Y chromosome is absent.

Plants

Plants display a great variety of sexual phenotypes. In particular, they show three possible options to sexually reproduce: i) to relegate the two sexes to separate individuals, ii) to keep them together on the same individual, iii) to have a combination of both (Tanurdzic & Banks, 2004). Genetic factors or environmental conditions control sex determination in land plants (Charlesworth, 2013). The majority of flowering plants are 'sexually monomorphic' species (Irish & Nelson, 1989, Charlesworth, 2002). This group is mainly

represented by hermaphrodite species, individual plants developing flowers that contain both pistils and stamens, or by monoecious species where the same individual produces separate male and female flowers. The 'sexually polymorphic' species are the minority in the plant kingdom. In this group the dioecious system, with separate male and female individuals, is found in only 9-10% of angiosperm species (Irish & Nelson, 1989, Charlesworth, 2002, Ming *et al.*, 2011).

The genetic of sex determination in plants involves, as in animals, two heteromorphic sex chromosomes, XY, with generally the male being the heterogametic sex. However, there are examples exhibiting the WZ system where the heterogametic sex is the female (Ming *et al.*, 2011). This diversified scenario includes also species in which sex determination is controlled by the X:A ratio (Matsunaga & Kawano, 2001).

Sex determination is traditionally considered to be the selective abortion or loss of function of male and/or female organs in the initially hermaphroditic floral primordia, resulting in unisexual flowers (Irish & Nelson, 1989, Charlesworth, 2002, Ming *et al.*, 2011).

Monoecy (individuals form unisexual male and female flowers, often physically separated, on the same individual), gynodioecy (dimorphic breeding system in which male sterile individuals (i.e., females) coexist with hermaphroditic individuals in populations), androdioecy (dimorphic breeding system in which female sterile individuals (i.e., males) coexist with hermaphroditic individuals in populations), and dioecy evolved from hermaphroditic ancestor species consequently to mutations in the genes involved in flower development causing male or female sterility (Irish & Nelson, 1989, Charlesworth, 2002, Ming *et al.*, 2011). To establish dioecy two genetic changes are required; one mutation aborting stamens (male sterile) and the other aborting carpels (female sterile). For example, in the XY (male heterogametic) system a recessive mutation (M => m) of the stamen-promoting-factor (SPF) on the homozygous XX chromosomes determines the male sterility and thus the female status. The Y chromosome contains a functioning male fertility allele as well as a dominant mutation (f => Su^F) on the gynoecium-suppressor-factor (GSF) at a

different locus that suppresses the development of female sex organs, and leads to the development of a male individual (Charlesworth, 2002, Ming *et al.*, 2011, Charlesworth, 2013). The regulatory pathways that have been modified during evolution from the hermaphrodite ancestors and to the emergence of dioecious species still remain largely unexplored.

Fungi

In contrast to animals and plants, fungal cell-type identity and sexual cycle are orchestrated by a more restricted chromosomal region, known as the mating type (MAT) locus. However, in several cases, clear parallels can be drawn between the structure of MAT and that of animal sex chromosomes (Fraser & Heitman, 2005). While the vast majority of sexually reproducing organisms occur as just two sexes or mating types, transitions from two to multiple mating types (MTs), and *vice versa*, have occurred in the fungal kingdom (Fraser *et al.*, 2007).

Sexual reproduction is common in fungi, and mating types occur in two general patterns: bipolar and tetrapolar (Fraser *et al.*, 2004). In the bipolar systems, a single genetic locus occurs in two alternative forms, known as idiomorphs (a or α, a or A, + or -, P or M) and these govern the identity of the cell (Metin *et al.*, 2010). Species with bipolar mating systems are found in the Ascomycete, Basidiomycete and Zygomycete phyla (e.g. *Saccharomyces cerevisiae, Schizosaccharomyces pombe,* and *Cryptococcus neoformans*) (Fraser *et al.,* 2004). Generally, two strains of different mating types, designated minus (-) or plus (+) are needed for successful mating. This leads to the formation of a zygospore, in which karyogamy occurs, followed by meiosis and the mitotic amplification of the progeny in a germsporangium structure to produce haploid germspores. Each mating type contains a unique gene, *sexM* or *sexP* (both encoding for a HMG-domain transcription factor), at the same position within the genome and flanked by the genes *tptA* and *rnhA*, which encode a predicted triose phosphate transporter and RNA helicase, respectively (Idnurm, 2011).

Basidiomycete fungi usually have more complex tetrapolar mating system, in which two unlinked genomic regions establish cell identity, and both must differ in two organisms involved in sexual reproduction (Fraser *et al.*, 2004). The maize pathogen *Ustilago maydis* is a tetrapolar basidiomycete with multiple mating types conferred by two mating type loci. One locus encodes pheromones and pheromone receptors, while the second encodes homeodomain transcription factors.

In contrast to the relatively small ascomycete MAT loci, the *C. neoformans* MAT loci are unusually large (spanning over 100 kb) and contain more than 20 genes (Fraser *et al.*, 2004) (Fig. 1.2). These MAT loci are regions of the genomes that exhibit similarities with sex-determining regions in other eukaryotes, including the presence of transcription factors, and dissimilar DNA regions between the alleles of each mating type (Idnurm, 2011).

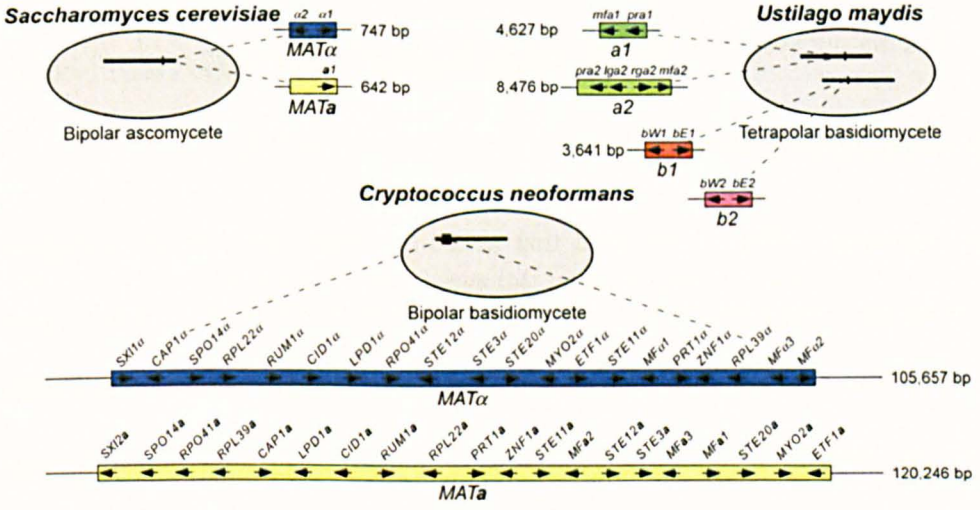


Figure 1.2: Fungal MAT locus in bi-polar and tetra-polar fungi (Fraser *et al.*, 2004).

Ciliates: *Tetrahymena thermophila*

The unicellular ciliate *Tetrahymena thermophila* has seven mating types. Cells can mate only when they recognize cells of a different mating type. *Tetrahymena* separates its germline and soma into two nuclei. During growth, the somatic nucleus is responsible for

all gene transcription while the germline nucleus remains silent. During mating, a new somatic nucleus is differentiated from a germline nucleus and mating type is decided by a stochastic process. In Cervantes *et al.* (2013), it is reported that the somatic mating type locus contains a pair of genes arranged head-to-head. Each gene encodes a mating type-specific segment and a transmembrane domain that is shared by all mating types. Somatic gene knockouts showed that both genes are required for efficient non-self-recognition and successful mating.

The germline mating type locus consists of a tandem array of incomplete gene pairs representing each potential mating type. Two classes of germline MAT alleles are known; the mat-1-like alleles encode mating types I, II, III, V, and VI, while mat-2-like alleles encode mating types II, III, IV, V, VI, and VII. During mating, a complete new gene pair is assembled at the somatic mating type locus; the incomplete genes of one gene pair are completed by joining to gene segments at each end of the germline array. All other germline gene pairs are deleted in the process. These programmed DNA rearrangements make ciliates a fascinating system of mating type determination (Cervantes *et al.*, 2013).

Amoebae: *Dictyostelium discoideum*

Urushihara & Muramoto (2006) discovered that sexually mature cells of *Dictyostelium discoideum* during gametogenesis present an overexpression (>100-fold) of RacF2 gene that encodes for a Rho GTPase resulting gamete-enriched. Through gene knockout and overexpression, the Authors isolated mutants showing anomalies in the extent of sexual cell fusion and asexual development, and suggested that RacF2 controls the process of sexual and asexual development through the regulation of cellular adhesiveness (Muramoto & Urushihara, 2006). However, it was Bloomfield *et al.* (2010) who discovered and analysed the mating-type locus of the model organism *Dictyostelium discoideum*. Three forms of a single genetic locus specify the three mating types of this social amoeba: two versions of the locus are entirely different in sequence, and the third

resembles a composite of the other two. Type I strains are characterized by a single protein-coding gene, matA, which is homologous to matB, one of the three genes present in the type II version of the locus. The two other genes making up the type II locus, matC and matD, are homologous to the two genes that are present in the type III version, matS and matT (Fig. 1.3) (Bloomfield *et al.*, 2010).

These results suggest a simple underlying picture: type I and type III mating behaviour can be specified by a single gene in each case: matA specifies type I and matS specifies type III. Type II is a composite in which homologs of matA and matS (matB and matC, respectively) allow it to mate with the other two mating types but, for reasons that remain unclear, not with itself (Bloomfield *et al.* 2010). The molecular function of these genes remains to be addressed.

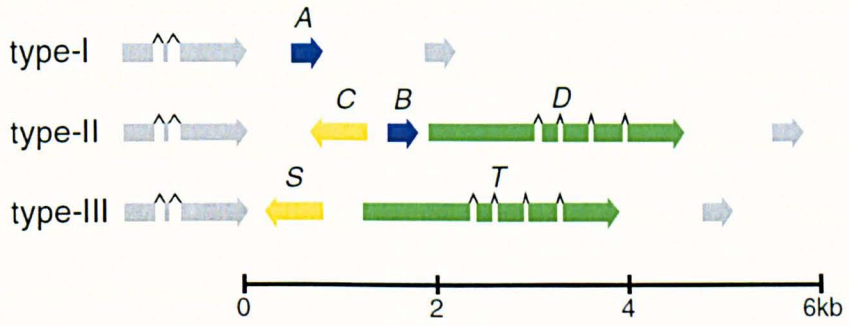


Figure 1.3: Structure of the *D. discoideum* mat locus (Bloomfield *et al.* 2010).

Algae

In algae investigations on the SD (sex determination)/MTD (mating type determination) system are very recent, but there are interesting examples to report.

In some algae, it is present a haploid phase determination system (UV system). The gametophyte, major stage of the life cycle occurs as separate male or female individuals that produce male and female gametes, respectively. Fertilization results in the UV non-sexed diploid phase (the sporophyte) and, when meiosis occurs, the sex chromosomes known as U and V assort in spores that carry either the U chromosome and give rise to

female gametophytes, or the V chromosome and give rise to male gametophytes (Bachtrog *et al.*, 2011).

In brown algae sex determination and sex-biased genes have been studied in *Ectocarpus siliculosus* and *Fucus vesiculosus*. For *E. siliculosus*, (Coelho *et al.*, 2011) and (Ahmed *et al.*, 2014) reported the identification (by linkage mapping) and the genetic and genomic characterisation of the U and V sex determining regions (SDR) of the algal model (by DNA and RNA deep sequencing). They found that sex was determined during the haploid phase by a non-recombining region of almost 1 Mbp. The SDR constituted only a fifth of the sex chromosome with a low number of sex-biased genes. The male and female haplotypes of the SDR were of similar size but were highly divergent; the only significant similarity was the presence of 11 homologs. Both haplotypes were rich in transposable element sequences and gene poor as compared to the autosomes, features typical of non-recombining regions. The male SDR haplotype was dominant over the female haplotype, suggesting that the V chromosome determines maleness, with femaleness possibly being the default state when the V chromosome is absent. A male-specific high mobility group (HMG) domain gene was identified as a candidate male sex-determining gene. This family of proteins is implicated in sex or mating type determination in both vertebrates and fungi (Bachtrog *et al.*, 2014). The SDR of the green alga *Volvox* also contains a HMG gene (Ferris *et al.*, 2010). In addition, the homologs were predicted to encode potential signal transduction proteins and could potentially be involved in the regulation of sex determination, while some of the male specific SDR genes were supposed to have a role in fertility.

The study of the fucoid brown alga *F. vesiculosus* provided the first transcriptomic analysis of expression variation in reproductive tissues for a brown alga during natural reproductive cycles (Martins *et al.*, 2013). The comparative analysis gained interesting information on the male and female sex-biased genes that regulate sexual reproduction. Results showed

that primary energy and carbohydrate metabolic pathways are under-represented in sexual (male and female) tissues. Differentiation is most clearly apparent in male tissue. At the same time, pathways for genetic information processing and cell-cycle related processes were over-represented in males. Moreover, specific transcripts were found in male receptacles that were not detected in females, consistent with sperm-specific developmental and signalling pathways, such as mitogen activated protein kinase (MAP2K), a cAMP-dependent protein kinase regulator (PRKAR), a PAS-PAC histidine kinase and putative blue-light photoreceptor, calmodulin genes and a Ca2+/calmodulin-dependent protein kinase (CaMK). This provides a general picture of male tissue as very active in signalling for a potentially diverse range of cellular processes (Martins *et al.* 2013).

In 2009 a tentative amplified fragment length polymorphism–simple sequence repeat (AFLP–SSR) linkage map of *Saccharina japonica* was constructed using a haploid population of 40 gametophyte clones leading to preliminary identification of the sex-determining region (Yang *et al.*, 2009). In the following years, a high density SNP linkage map was constructed for the same species. The RAD tags for a gametophyte clone mapping panel were also extended so that a SNP chip could be developed. In addition, a set of microsatellites were identified among mapped loci, and a gametophyte sex determining locus was mapped (Zhang *et al.*, 2015). However, no genomic information is still available for this species.

In the Volvocine algae, that are a group of chlorophytes comprising unicellular species such as *Chlamydomonas reinhardtii* and multicellular species such as *Volvox carteri*, a sexual cycle has been recognised and the MT-locus much more deeply studied with respect to other algae. Remarkable expansion and divergence relative to the MT locus are present between C*hlamydomonas* and *Volvox*. The first one undergoes a sexual cycle, regulated by environmental conditions and by cell–cell interactions, in which a large haploid SDR of

~1Mb controls sexual differentiation, mating compatibility, and zygote development triggered by nitrogen deprivation. The mating locus is a multigenic chromosomal region within which gene order is rearranged in the two sexes (MT+ and MT–) and meiotic recombination is suppressed, thus leading to its inheritance as a single Mendelian trait. Within each MT-locus there are sex specific genes , which are required for the sexual phase, as well as shared genes present in both sexes, most of which have unknown function (Pan & Snell, 2000, Goodenough *et al.*, 2007, Ferris *et al.*, 2010). In *Chlamydomonas,* sex-related genes are both MT-biased and autosomal. The MT+ and MT− loci region is characterized by several large inversions and translocations, presumably contributing to recombinational suppression. The SDR presents a central rearranged (R) domain flanked by centromere-proximal (C) and telomere-proximal (T) sequences, which also fail to recombine. The MT+ R domain contains three DNA regions not found in the MT− locus, as well as a block of two tandem genes: *EZY2*, whose expression is confined to the zygote, alternating with *OTU2* that is expressed exclusively in MT+ gametes (Goodenough *et al.*, 2007). Reciprocally, the MT− locus contains three regions not found in the MT+ locus. Relevant genes resident in these regions are *FUS1*, *MTD1* and *MID,* to whom have been assigned MT-specific functions in gametogenesis and mating. Two genes, *MID* and *MTD1*, were directly involved in activating MT- gametogenesis. The *MID* gene, unique to region of the MT-locus, is so-named because it is responsible for the MT- dominance; cells expressing a *MID* gene differentiate as *minus.* In particular, *MID* represses the autosomal gene encoding the MT+ agglutinin glycoprotein (*SAG1*), and activates the MT- agglutinin gene (*SAD1*); so *MID* is necessary both to activate *minus* gene expression and to prevent *plus* gene expression. The MID protein is a bZIP transcription factor in the RWP-RK family (Goodenough *et al.*, 2007).

In *Chlamydomonas,* the specific adhesion between gametes of opposite mating type generates signalling pathways that quickly render the gametes refractive to additional adhesive interactions and initiates zygote development. Moreover the cell–cell fusion event

occurs at plasma membrane sites and through plasma membrane molecules that are distinct from those responsible for the initial recognition/adhesion between the two types of gametes (Pan & Snell, 2000).

Chlamydomonas is isogamous (producing equal-sized gametes) while *Volvox* has evolved oogamy that is under the control of female and male MT loci. The sexual cycle of *Volvox* is characterized by a suite of other traits not found in *Chlamydomonas*, such as a diffusible sex-inducer protein rather than nitrogen deprivation (–N) as a trigger for gametogenesis (Ferris *et al.*, 2010). *Volvox* presents a MT locus ~500% larger than the *Chlamydomonas* one, containing over 70 protein-coding genes in each allele and a diffusible sex-inducer protein as a trigger for gametogenesis (Pan & Snell, 2000, Goodenough *et al.*, 2007, Ferris *et al.*, 2010). In *Volvox* only two genes from *Chlamydomonas, MID* and *MTD1*, had recognizable homologs. Both *MTD1* and *MID* were present in the male but expressed constitutively, indicating that their transcription was uncoupled from sexual differentiation (Ferris *et al.*, 2010). This result suggests that additional MT genes might play a role in gametogenesis. Ferris *et al.,* (2010) used differential deep transcriptome sequencing for the identification of new MT-limited genes. The transcriptome data provided a list of genes with a sex-biased expression and sex-regulated expression. This set of genes encode putative signalling, extracellular matrix, and chromatin associated proteins with known or potential roles in gametogenesis and fertilization. Two interesting female-biased genes were found. *FSI1* was strongly induced during gametogenesis encoding a small predicted transmembrane protein and *HMG1* was encoding a HMG domain protein that belongs to a family of DNA binding proteins whose members also regulate mammalian and fungal sex determination. It was the first time that HMG proteins were reported in the sex determination systems of green algae.

Also in the volvocine genus *Gonium MID* homologs were identified and their presence/absence was examined in nine strains of four species by Hamaji *et al.* (2013).

These isogamous species have a heterothallic mating system, with mating types designated arbitrarily as plus or minus, or a homothallic system. This study provided a framework to assign heterothallic mating types through the use of homologous molecular markers among lineages (Hamaji *et al.*, 2013).

Several studies have been carried out on sexual reproduction of the unicellular Charophycean *Closterium peracerosum-strogosum-littorale* Complex. The species is heterothallic and when mixing together the two haploid compatible mating types (MT+ and MT-) in nitrogen-depleted medium in the light, a particular sexual cell division (SCD) takes place. After pairing of the haploid vegetative cells, SDC produces haploid sexually competent gametangial cells and the formation of the conjugation papillae. Gametangial cells condense their cytoplasm, produce gametes and, following conjugation, form the zygospore (Charlesworth, 2002, Tsuchikane *et al.*, 2010). Zygospores acquire resistance to dry conditions and become resting stages. Exposure to dry conditions and subsequent water supply lead to the start of meiosis, resulting in a pair of MT+ and MT- cells arising from one zygospore (Sekimoto *et al.*, 2014). The process of sexual reproduction in *Closterium* is well characterized both physiologically and biochemically. The two major pheromones involved in the process and promoting multiple steps all along the conjugation phase are PR-IP (Protoplast Release-Inducing Protein) Inducer and PR-IP. Both are glycoproteins, the first one released constitutively from MT- directly induces the production and release of PR-IP from MT+, whereas PR-IP induces SCD with the release of mucilage and gametic protoplast from MT- cells. The release of protoplast in MT+ cells is probably induced by direct adhesion to MT- cells during pairing. However, it is still unknown what leads to cell-cell recognition and fusion; probably the process is triggered by a third chemotactic pheromone. The genes encoding the two pheromones are present in both mating types but they result differentially expressed (Sekimoto *et al.*, 2014).

The molecular mechanism underlying intercellular communication and sex determination system in *Closterium* has been investigated through EST (expressed sequence tag) (Sekimoto *et al.*, 2003) and microarray analyses (Sekimoto *et al.*, 2006). Two genes involved in sexual reproduction, *CpRLK1* and *CpRLP1*, resulted to be pheromone-inducible and conjugation-related. The first one encodes a receptor-like protein kinase containing an extracellular domain, a transmembrane domain, and a kinase domain. It is localized on the conjugation papilla of the MT+ and its knockdown impairs the release of the protoplast and zygote formation. *CpRLK1* is probably an ancient cell wall sensor that now works to regulate osmotic pressure for a proper protoplast release. *CpRLP1* encodes a receptor-like protein containing eight leucine-rich repeats (LRRs) in the extracellular domain, a single transmembrane domain, but no kinase domain. Its expression is promoted in MT- cells in response to PR-IP. It may form a heterodimer and it is also probably involved in protoplast release after pairing (Hirano *et al.*, 2015).

Diatoms

The knowledge on the molecular genetics underlying mating system determination in diatoms is still in its infancy.

Vanstechelman *et al.* (2013) provided the first evidence for a genetic sex determining mechanism in a benthic diatom, the model species *Seminavis robusta.* An AFLP-based strategy was employed to identify MT-linked AFLP markers. 13 MT+ and 15 MT- linkage groups were obtained from the analysis of 463 AFLP markers. Five linkage group pairs could be identified as putative homologues and the mating type phenotype mapped as a monogenic trait, disclosing the MT+ as the heterogametic sex (Vanstechelman *et al.*, 2013). Data were confirmed with BSA (Bulked Segregant Analysis). The genetic structure of the MT locus was identified as a SF2-family related Helicase/S adenosyl methyltransferase (HEL-SAM) with a transcript length of 7866 bp (W. Vyverman, personal communication). During sexual reproduction it was also identified the production

of a sex pheromone. MT- cells probably produce a primary signal that activates MT+ cells. These cells start secreting a sex-inducing pheromone responsible of the light-dependent production of L-diproline by MT- gametangia. This pheromone was capable of attracting MT+ gamentagia (Gillard *et al.*, 2013, Moeys *et al.*, 2016).

1.3.2 The variety of sex determination systems and primary sex determining genes

To understand the process of sex determination, we have to consider the different mechanisms that have been uncovered and the evolution of the sex-biased genes.

Dual sex chromosome systems, in which either the female (ZW/ZZ) or the male (XX/XY) is heterogametic, are common above all in vertebrates and plants. Other systems, as in invertebrates like *Drosophila melanogaster and Caenorhabditis elegans*, are set by the ratio of the number of X chromosomes to sets of autosomes (X:A) (Haag & Doty, 2005) and, finally, an haploid phase determination system (UV system) as in some algae and bryophytes. In fungi mating types are set by two alternative forms of a single genetic locus (bipolar systems) or by two unlinked genomic regions (tetrapolar system).

In mammals, sex determination depends upon the primary sex-determining gene *SRY* (Sex determining region Y), while in invertebrates, such as *Drosophila melanogaster* and *Caenorhabditis elegans*, sex determination depends respectively upon the key sex-determining genes *doublesexd (dsx)* and *mab-3,* both genes encode proteins with a DNA-binding motif (DM domain) (Raymond *et al.*, 1998, Haag & Doty, 2005). No master sex determination gene has been identified in dioecious plants (Bachtrog *et al.*, 2014). In fungi the sex determined genes is represented by a HMG-domain transcription factor (Idnurm, 2011).

The discovery of the homology of *dsx* in *Drosophila melanogaster* and *mab-3* in *C. elegans* provided the first evidence for a common evolutionary basis of sex determination in animals (Bachtrog *et al.*, 2014). The majority of the species have been screened for sex-

specificity of genes related to *doublesex-mab-3* (DM)-family genes with roles in male sexual development *SRY*, whose important role has been well established. These loci have been identified as the primary sex-determining transcription factor genes in the medaka fish (*Oryzias latipes*), in most mammals and in insects. The DM domain is described as a zinc finger-like DNA binding motif and *SRY* encodes a DNA binding protein of the HMG-box (High Mobility Group box) family that recognizes both chromatin structure and a specific binding sequence. Genes related to the *doublesex-mab-3* (DM)-family, which play a role in male sexual development, were discovered in vertebrates and even cnidarians. For example, in humans *DMRT 1* (*doublesex* and *mab-3* related transcription factor) belongs to the family of genes that encode proteins containing DM-domain, a novel DNA-binding motif. *DMRT 1* is one of the most conserved genes in sex determination, since its presence has been observed across phyla, from invertebrates to vertebrates (Haag & Doty, 2005, Rai & Roy, 2008). Compared to the diversity of the mode of sex determination and the identity of the master-switch genes, some key regulatory genes play conserved roles in the molecular pathways leading to male or female gonad development across invertebrates and vertebrates, such as the *doublesex-mab3* (DM) family genes (Bachtrog *et al.*, 2014).

The current knowledge on sex determination mechanisms, primary sex determining gene and sex-biased ones in algae is summarized in Table 1.2.

Table 1.2: Known sex determination mechanisms, master sex-determining genes and sex-biased genes in algae. Master sex-determining genes are indicated as (demonstrated) or (candidate) whether or not their role was confirmed by experimental validation.

Species	Sex-determining mechanisms	Primary sex-determining gene	Sex-biased genes	Sex-biased genes function	References
Chlamydomonas reinhardtii	Haploid MT	MID (bZIP TF) (demonstrated)	*SAG*), *SAD1*, *FUS1*, *Gsm1/Gsp 1*	gametogenesis and fertilization competences, zygote differentiation, and control of organelle inheritance	Ferris *et al.*, (2010), Goodenough *et al.*, (2007)

Volvox carteri	Haploid MT	HMG-domain TF (candidate)	Many male-biased genes and female genes as *FSI1,HMG1*	extracellular signalling, gametogenesis and fertilization	Ferris *et al.*, (2010)
***Closterium peracerosum-strogosum-littorale* complex**	Haploid MT	Unknown	*CpRLK1, CpRLP1*	probably involved in protoplast release process	Sekimoto *et al.*, (2006); Hirano *et al.*, (2015)
Ectocarpus siliculosus	Haploid UV	HMG-domain TF (candidate)	various	microtubule and calcium binding-related processes and photosynthesis	Coelho *et al.* (2011); Ahmed *et al.* (2014); Lipinska *et al.*, (2015)
Fucus vesiculosus	Diploid MT	Unknown	various	sperm-specific developmental and signalling pathways	Martins *et al.* (2013)
Seminavis robusta	Diploid MT	Unknown	Unknown	Unknown	Vanstechelman *et al.*, (2013)
Pseudo-nitzschia multistriata	Diploid MT	Unknown	*MRP1, MRP2, MRM1, MRM2*	Transcription factors and signalling pathways	Vitale PhD thesis

1.3.3 Why and how to study sex determination

A) Why to study sex determination

There are a series of motivations to approach the study of sex determination systems in diatoms.

Explore the level of conservation of sex determination systems within diatoms. The model organism of this PhD project is a pennate (class Bacillariophyceae Haeckel) diatom. The vast majority of species within this lineage are characterized by heterothallic mating systems with two mating types, MT+ and MT-. This is also the case of *Seminavis robusta*, a benthic pennate diatom for which a single sex locus was identified through an AFLP-based sex-specific linkage map approach (Vanstechelman *et al.*, 2013). The two model diatoms have different habits, i.e. *P. multistriata* is marine and planktonic, whereas *S. robusta* is a brackish and benthic species. The two species also belong to different families:

P. multistriata to Bacillariaceae and *S. robusta* to Naviculaceae. One motivation of this study was to test the level of conservation of the sex determining locus and sex-related genes between these two pennate diatoms. This will constitute the basis for exploring evolutionary comparative analyses of sex determination systems and sex-related genes amongst the diversity of diatoms through the analysis of the available diatom genomes. It is considered that the evolution of separate mating types/sexes is an evolutionary derived feature, found primarily among multicellular organisms (Bachtrog *et al.*, 2014). However, there are exceptions and heterothallic pennate diatoms are one of them.

Explore sex ratios in natural populations. Pseudo-nitzschia multistriata has a heterothallic mating system, but cells of opposite mating type are morphologically not distinguishable. The mating type of a considerable number of clonal strains of this species, isolated in year 2008, 2009 and 2010, has been assessed *via* crossing experiments with pairs of strains of known mating type (Scalco, 2013; for details of the method see Chapter 2). Interestingly, 92.2% of the strains isolated in 2008 turned out to belong to MT-, while percentages were more balanced for strains isolated in 2009 (50.8% MT+ and 49.2% MT-) and 2010 (37.9% MT+ and 56.1% MT-). These very puzzling results raise questions about the mechanisms that determine mating types in this diatom. The relatively balanced mating type ratio recorded in 2009 and 2010 would support a genetic sex determination system, as in the benthic pennate diatom *Seminavis robusta* (Vanstechelman *et al.*, 2013). If sexes/mating types are determined by the presence or absence of an allele on a single gene locus, the random segregation of genes at meiosis will produce a balanced sex ratio. Unbalanced sex ratios have been reported in some organisms, e.g. lizards, as the result of the interplay between genotypic sex determination and environmental sex determination (Uller *et al.*, 2007). In the brown algae *Laminaria saccharina* and *L. religiosa* it has been shown that sex ratio can be modified by environmental stressors such as salinity or temperature (Bartsch *et al.*, 2008). Sex ratio seems to be mainly genetically determined in two

populations of the brown alga *Lessonia nigrescens*, but temperature can significantly modify it (Oppliger *et al.*, 2011).

Unbalanced sex ratios can have profound implications for population dynamics of microalgae, although no information is presently available – besides the unpublished information reported above – on the distribution of mating types of unicellular organisms. To address these questions, we need to know the genetic architecture of the sex locus of *P. multistriata* in order to design sex markers, an approach that has been recently set up for kelps (Lipinska *et al.*, 2015).

Explore the role of sex-biased genes. Besides the possible identification of the sex-determining locus of *P. multistriata,* I expect that the comparative transcriptomic approach I plan to use will provide a series of sex-biased genes. Sex-biased genes have a differential expression in individuals of opposite sexes/mating types and determine a considerable number of developmental, morphological and physiological characters in multicellular organisms (Ellegren & Parsch, 2007, Parsch & Ellegren, 2013). The heterothallic diatom *P. multistriata* is a good candidate species for the exploration of sex-biased genes in diatoms. Expected functions of these genes should be related to e.g. mate recognition, a phase in the life cycle in which cells of opposite mating type perform in a different way. Complex multi-phasic interactions between cells of opposite mating types mediated by sex pheromones have been reported for the diatoms *Seminavis robusta* (Gillard *et al.*, 2013; Moeyes *et al.*,2016) and *Pseudostaurosira trainorii* (Sato *et al.*, 2011) and there is evidence that chemical cues are effective also in *P. multistriata* (Scalco *et al.*, 2014). Genes involved in mating type specific production and perception of pheromones are thus possible candidate sex-biased genes. These genes are expected to play a role in pre-zygotic reproductive barriers, allowing mating only between conspecific cells, and could be potential candidates for studying speciation processes (Coyne & Orr, 2004).

B) How to study sex determination

Two approaches have been largely used to identify the sex locus in a broad range of organisms: genetic and genomic. The examples illustrated in this chapter span from the oldest AFLP technique for linkage mapping analysis to the differential expression analysis to find sex (MT)-biased genes. Both genetic and genomic approaches are essential to achieve the goal and one does not exclude the other. This assumption is supported by the evidence that many comparative transcriptional analyses led to the discovery of sex-biased genes yet without identifying the sex locus (e.g. *Fucus vesiculosus* (Martins *et al.*, 2013), unless they were combined with genetic analysis, as in *Ectocarpus siliculosus* (Coelho *et al.*, 2011, Ahmed *et al.*, 2014). In the ciliate *Tetrahymena thermofila,* the molecular identification of the mating type locus was obtained using RNA-Seq, but a genetically mapped region of about 300 Kb was previously identified as *mat* locus by the analysis of meiotic recombination frequency on a linkage group. Again the combination of genetic and genomic approaches resulted successful. In the unicellular algal species *Chlamydomonas reinhardtii,* the MT- locus was identified in the early nineties (Ferris & Goodenough, 1994, Goodenough *et al.*, 1995) when NGS approaches were not available, and the yeast *MAT* locus was identified even much earlier (Klar, 2010).

1.4 Molecular tools for diatoms

It has been only in the last 11 years that complete genome sequences of diatoms became available. Up to now, we have access to five genome sequences of both centric and pennate diatoms representative species:

Thalassiosira pseudonana (Armbrust *et al.*, 2004)

Thalassiosira oceanica (Lommer *et al.*, 2012)

Phaeodactylum tricornutum (Bowler *et al.*, 2008)

Pseudo-nitzschia multiseries (http://genome.jgi-sf.org/Psemu1/Psemu1.home.html)

Fragilariopsis cylindrus (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html)

Also the genome of *Fistulifera solaris* has been recently sequenced; however, it is not yet accessible (Tanaka *et al.*, 2015). *Thalassiosira pseudonana* and *T. oceanica* represent the group of centric diatoms, while the other three are pennate diatoms. The genomes of *P. tricornutum* and *T. pseudonana* are of 27.4 megabases and 34.5 megabases, respectively, with 10,000 and 14,000 genes, respectively. The sequenced genomes had only half of the genes with an assigned function while ~35% of the genes were reported to be species-specific. In diatom genomes certain gene families are expanded as compared to other eukaryotes. For instance, this is the case for the cyclins family (Huysman *et al.*, 2010) and the heat shock factor family of transcription factors, which amount to approximately 50% of the total number of transcription factors reported from *P. tricornutum* and *T. pseudonana* (Montsant *et al.*, 2007, Rayko *et al.*, 2010). These diatom specific gene family expansions could explain the adaptability of diatoms in rapidly changing environments and responses to various environmental signals such as availability of nutrients as well as biotic and abiotic stresses. However, since the prediction of diatom gene functions is around 55%, functional genomics and reverse genetics approaches to further explore diatom gene repertories are required.

Besides genomes, ESTs libraries and transcriptomes of different species have been produced as part of the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP), one of the massive effort where whole transcriptomes of over 650 marine micro eukaryotes have been generated using NGS technology (Keeling *et al.*, 2014). Among these latter, *Pseudo-nitzschia delicatissima, Pseudo-nitzschia arenysensis and Skeletonema marinoi,* for which we had fast access (even before their publication), were produced at SZN. Comparative transcriptomic analysis among *Pseudo-nitzschia delicatissima, Pseudo-nitzschia arenysensis* and *Pseudo-nitzschia multistriata* permitted to annotate about 80% of the sequences in each transcriptome and to compare the main metabolic pathways, finding out distinct species-specific patterns as also general pathways (e.g. urea cycle, C4 photosynthetic pathway, fatty acid oxidation) first thought to be exclusive to plants and animals (Di Dato *et al.*, 2015).

These advances greatly facilitate functional genomics research in many diatoms, organisms that, in spite of their tremendous ecological importance, have their molecular mechanisms largely unexplored (Bowler *et al.*, 2010, Haas *et al.*, 2013).

Novel tools to modulate gene expression, like overexpression and gene silencing, have been developed for the model species *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Siaut *et al.*, 2007, De Riso *et al.*, 2009, Bertrand *et al.*, 2012) and are in development also for *P. arenysensis* and *P. multistriata.* Sabatino *et al.* (2015) achieved the first genetic transformation of the planktonic diatoms *P. arenysensis* and *P. multistriata* with the biolistic method, using the H4 gene promoter from *P. multistriata* to drive expression of exogenous genes.

The synergic development of new molecular tools, the advent of the omics era and a more intensive study of diatoms in the last decades fostered an accumulation of assorted information. With the attempt to explain complex ecological scenarios and peculiar biological traits of single species we often rely on their genomes and transcriptomes,

however we still need to improve the molecular approaches necessary to decode and exploit this information (Chepurnov *et al.*, 2008, Sabatino *et al.*, 2015).

1.5 *Pseudo-nitzschia multistriata* as model organism for genomic studies

The genus *Pseudo-nitzschia* includes species of marine diatoms that can be responsible for blooms in both coastal waters and open oceans. *Pseudo-nitzschia* species are identified based on the presence/absence and combination of different morphological and ultrastructural characters: cell shape and width, density of striae and fibulae (number per 10 μm), morphology and density of perforations (areolae) (e.g.; (Lundholm *et al.*, 2003, Lundholm *et al.*, 2006, Amato & Montresor, 2008, Lundholm *et al.*, 2012)). Identification at the species level often requires detailed investigations in transmission electron microscopy. Studies on the genetic diversity of *Pseudo-nitzschia* species have been carried out using different molecular markers such as ITS, LSU, rbcL (e.g. (Lundholm *et al.*, 2003, Lundholm *et al.*, 2006, Amato, 2007, Amato & Montresor, 2008, Casteleyn *et al.*, 2008, Quijano-Scheggia *et al.*, 2009b, Lundholm *et al.*, 2012)).

Pseudo-nitzschia multistriata (Fig. 1.4) is one of the members of this genus and it has been recorded at the Long Term Station in the Gulf of Naples (Tyrrhenian Sea, Italy) since 1995. *Pseudo-nitzschia multistriata* (Takano) Takano is a chain forming planktonic, raphid pennate, diatom described from Japanese waters as *Nitzschia multistriata* (Takano, 1993) and subsequently transferred to the genus *Pseudo-nitzschia* (Takano, 1995). It blooms in summer and early autumn (D'Alelio *et al.*, 2010) and produces the neurotoxin domoic acid (Orsini *et al.*, 2002), a small amino acid that acts as an analogue of the neurotransmitter glutamic acid, causing Amnesic Shellfish Poisoning (ASP). In the Mediterranean Sea there were no reports of ASP intoxications up to now. *P. multistriata* can be recognised by light microscopy and distinguished from other *Pseudo-nitzschia* by its prominent sigmoid shape in girdle view; it can be easily cultivated and can be stimulated to reproduce sexually under controlled laboratory conditions (D'Alelio *et al.*, 2009).

Its life cycle is heterothallic (Fig. 1.5) comprising two opposite mating types (MT+ and MT-) (D'Alelio *et al.*, 2009). The life cycle of *P. multistriata* conforms to the general

pattern of other pennate planktonic species (Davidovich & Bates, 1998, Amato *et al.*, 2005, Chepurnov *et al.*, 2005). The life cycle is 'cis-type', in which one gametangium produces passive (-) gametes that remain attached to the empty gametangium and the other gametangium produces active (+) gametes that escape from gametangium and migrate toward the passive gametes to fuse (D'Alelio *et al.*, 2009, Scalco *et al.*, 2015). The fusion produces two zygotes that remain attached to the (-) gametangium and develop into an elongate auxospore within the large initial cell is produced.

As most diatoms, also *P. multistriata* undergoes a progressive reduction of the average cell size of its populations as consequence of vegetative growth (mitotic division); the formation of large-sized cells occurs within sexual reproduction meaning that sexuality is the only strategy to survive and avoid extinction. Sexual events can be easily induced when monoclonal cultures of opposite mating type, below the SST, are grown together. The sexualisation size threshold can span from 60-55 µm to 26 µm, measure of the smallest sexually active cells studied. The maximum cell size of initial cells was reported between 72 to 81 µm (D'Alelio *et al.*, 2009).

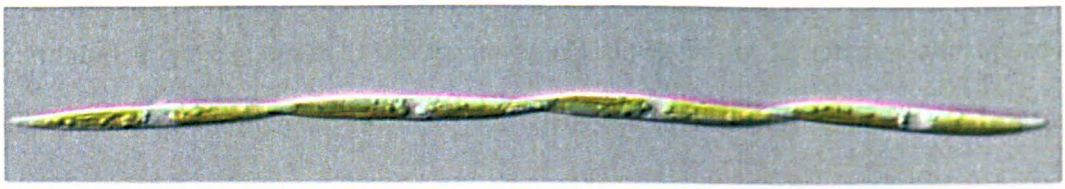


Figure 1.4: Photograph of cells in chain of *P. multistriata*.

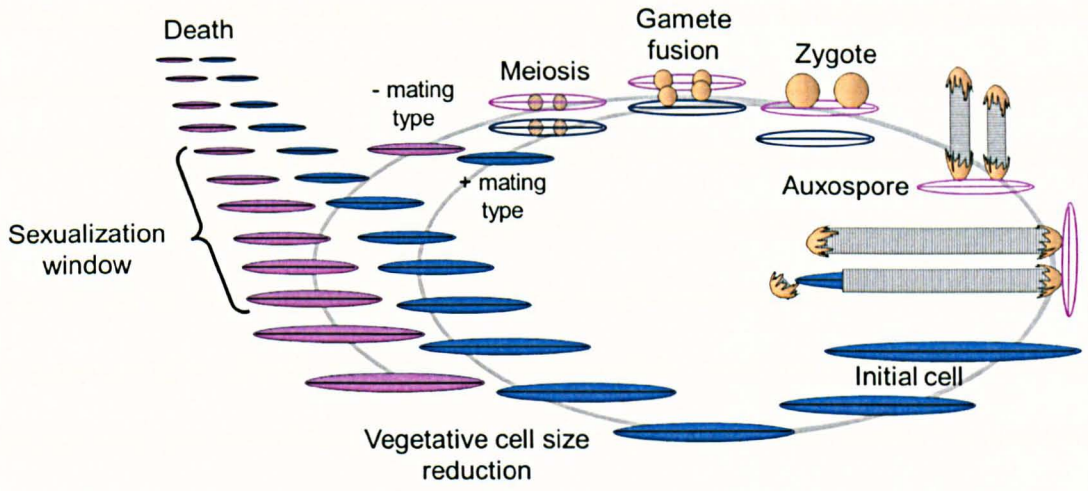


Figure 1.5: scheme of the life cycle of a *Pseudo-nitzschia* species (pennate diatom) (Scalco *et al.*, 2015).

Few studies have addressed the implication of sexual reproduction on population dynamics (D'Alelio *et al.*, 2010). The Authors monitored the cell abundances and cell size patterns of cells in natural samples collected in the Gulf of Naples over 10 years and the implementation of an individual-based model allow to infer the biennial occurrence of sexual reproduction (D'Alelio *et al.*, 2010). In the natural environment large cells produced following a sexual event were found every two years; however, the maximum sized initial cells were never recorded probably due to their low concentration. Based on the parameters determined from laboratory cultures merged with the information from natural populations, the authors developed a model of population growth in which sex occurs in year 1 and produces a small fraction of initial cells; large cells become more abundant and thus detectable in year 2 during which these cells progressively decrease their average size during the bloom periods and reach the cell size window for sex. Two years after the predicted sex event, two cell sizes are present in the population: the mature ones able to start a new sex cycle and a cohort of rather small cells that presumably cannot undergo sex. This life cycle framework supports the occurrence of sex in alternate years (D'Alelio *et al.*, 2010).

The size of the population is not the only parameter that influences sexual reproduction. The importance of signalling in phytoplankton is well known and several molecules have been suggested to act as infochemicals at sea. There are evidences that a signalling process takes place during *P. multistriata* sexual reproduction (Patil, 2014, Scalco *et al.*, 2014); although very little information is available on the type of molecules that act during early stages of mating and on how cells receive and process signals. The analysis of microsatellite patterns in F1 cells produced by different parental strains provided the genetic proof of sexual reproduction, showing that microsatellites are inherited with a Mendelian pattern (Tesson *et al.*, 2013). Finally Patil *et al.* (2015) identified a meiotic toolkit of 42 genes potentially involved in meiosis shared between *P. multistriata* and other five diatom species. A transcriptomic approach was used to analyze the expression rates of the transcripts belonging to the meiotic toolkit and for 37 of them the expression levels resulted higher during meiosis when compared to the vegetatively growing monoclonal cultures validating their meiotic role, while phylogenetic analyses revealed a recent expansion in the RAD51 family in diatoms.

Availability of whole genome sequence, small genome size, RNA-seq data, rapid growth properties, easiness to culture and a well-described and controllable life cycle are some of the attributes that make a species a model organism. *Pseudo-nitzschia multistriata* possess all these attributes and compared to the most widely used diatom models, that are apparently asexual, *P. multistriata* is able to undergo sexual reproduction allowing investigations on the molecular mechanisms regulating different aspects of the sexual phase.

The genome of *Pseudo-nitzschia multistriata* has been sequenced at The Genome Analysis Centre (TGAC) in Norwich (UK) (http://www.tgac.bbsrc.ac.uk/ccc/) using the Illumina/ Solexa sequencing technology, within a collaborative project funded by the Stazione Zoologica Anton Dohrn coordinated by Dr. Maria Immacolata Ferrante. The sequenced

clonal strain is a MT+ (B856) from the second generation in the pedigree of *Pseudo-nitzschia multistriata* (Fig. 1.6). The strain was generated by crossing two sibling strains from the F1 generation, thereby reducing heterogeneity in the genome. The sequencing and downstream assembly yielded a genome of 59 MB composed of ~1000 scaffolds with an N50 of 139 Kb. Approximately 12,000 genes are predicted in the assembled scaffolds and annotated with the Annocript pipeline, where ~80 % genes were assigned a Uniprot ID (9653 genes) and additional 214 genes were exclusively annotated on the basis of a CDD domain profile (Basu *et al.*, under revision). The genome is accessible through a genome browser that will be soon public, with different available tracks, i.e. RNA-seq raw reads, transcriptome, gene model prediction, restriction sites, etc., and blast tool.

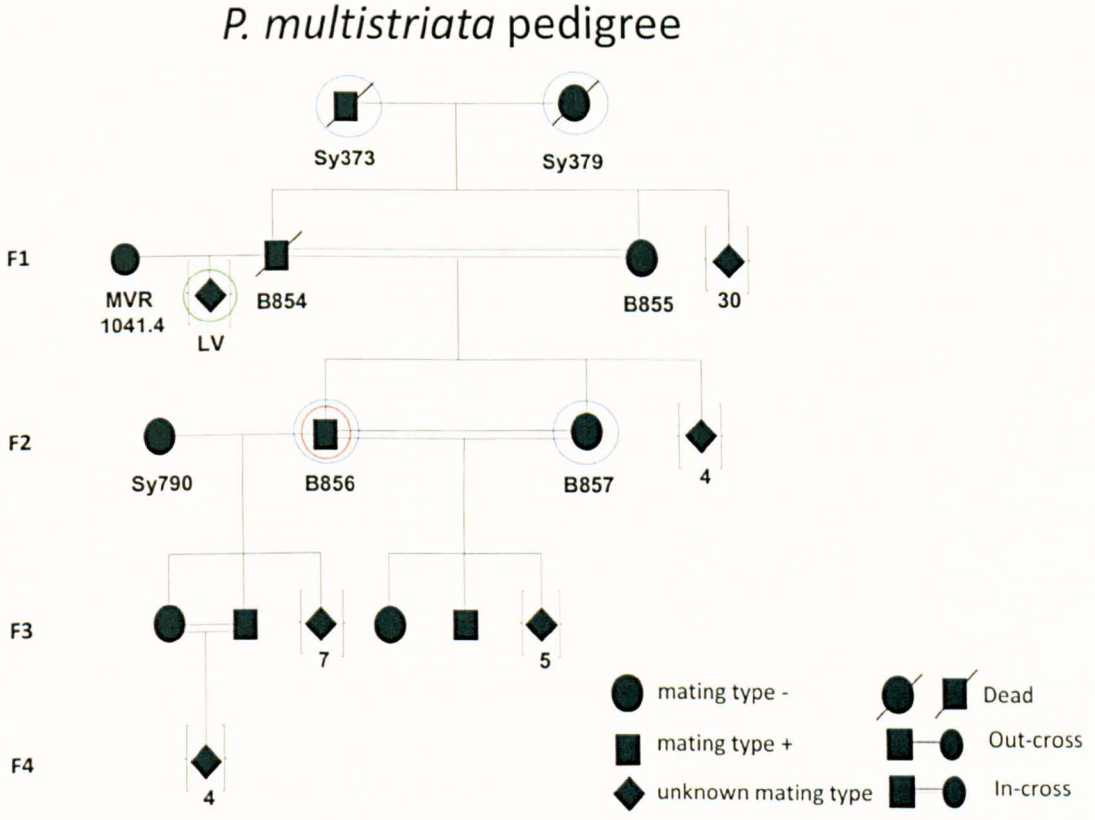


Figure 1.6: *P. multistriata* pedigree. Pedigree of *Pseudo-nitzschia multistriata* consisting of clonal strains from four consecutive generations. The strains SY373, SY379, B856 and B857 have been used for RNA-seq (circled in blu), while strain B856 has been used also for genome sequencing (circled in red). The LV strains are the ones that will be used as mapping population (circled in green).

The *de novo* transcriptome of *P. multistriata* has been sequenced by the Joint Genome Institute (JGI) within the project 'A deep transcriptomic and genomic investigation of diatom life cycle regulation' funded by the same JGI using Illumina HighSeq on six libraries (three MT+ and three MT-). RNA-seq has been performed also on other 16 libraries of *P. multistriata* during early stages of sexual reproduction (Patil, 2014). The results of these transcriptomic studies will be further discussed in the Chapter 2 of this thesis.

It is important to outline the chronology of production of the molecular tools for *P. multistriata* through the years. The sequencing of the first six RNA-seq libraries started in 2011 and ended in 2012. The first *de novo* transcriptome assemblies were produced by JGI between 2011 and 2013. The following versions were improved by Dr. Remo Sanges (Stazione Zoologica Anton Dohrn) until the final assembly was created at the end of 2013 (Chapter 2). The genome of *P. multistriata* was sequenced in 2012 and four assemblies were produced along 2012-2014 to finally achieve the last version (V1.4) available on the genome browser. At the beginning of 2014 the 'sensing transcriptome' was produced and analysed (Basu *et al.*, under revision, Patil, 2014, and Chapter 3 of this thesis). Finally, in 2015 the transformation of *P. multistriata* was set up (Sabatino *et al.*, 2015) and the genome of two MT- and three MT+ were re-sequenced. Their mapping on the reference genome is still on-going (Table 1.3).

Table 1.3: Chronology of the production of *P. multistriata* molecular tools.

	START	END
P.multistriata de novo transcriptome assembly	2011	2013
P.multistriata de novo genome assembly	2012	2014
P.multistriata 'sensing transcriptome'	2014	2014
P.multistriata transformation	2014	2015

P.multistriata genome re-sequencing	2015	-

1.6 Aims of the thesis

The general aim of my PhD project was to investigate the molecular bases of sex determination system of marine diatoms. The model species is the marine planktonic raphid diatom *Pseudo-nitzschia multistriata,* for which several genomic tools are and will be available. *Pseudo-nitzschia multistriata* has a heterothallic mating system, i.e. sexual reproduction can be induced only when strains of opposite mating type (MT) get in contact. The specific objectives of this thesis were the identification of the MT locus and of mating type-related genes.

The persistence of a diatom population in a certain area depends on the frequency of sexual events producing large-sized cells; if sex does not occur, the population risks extinction. An unbalanced sex ratio in natural populations can impair the frequency of sexual reproduction. The identification of mating-type locus and MT-biased genes will thus allow to apply demographic approaches study of natural diatom populations. Moreover, MT-biased genes are involved in various processes related to sexual reproduction, such as synthesis and perception of pheromones, signalling molecules, systems related to gamete-gamete recognition and conjugation. Elucidating the molecular and genetic bases of MT determination and sex-biased genes will thus also contribute to understand mechanisms of reproductive isolation, speciation, evolution of life cycles, and establish the diatoms as a novel model group to study the evolution of reproductive strategies in eukaryotes.

The questions I have addressed are:

Which genes are differentially expressed in the two MTs of *P. multistriata*? Could they be candidate MT-determining genes? How many genes are mating type-biased?

Chapter 2 illustrates the differential expression analysis conducted on the RNA-seq dataset of 3 MT+ and 3 MT- strains of *P. multistriata* within the sexualisation cell size range. This is the first transcriptomic analysis of differential gene expression between opposite mating

types in diatoms. The analysis allowed the identification of five MT-biased genes, validated by qRT-PCR. It follows a detailed description of *P. multistriata* MT-biased genes. A computational characterization was conducted to understand their putative function in sexual reproduction. The MT-biased genes were further analysed to study the selective pressure acting on them performing a Ka/Ks calculation.

Are the MT-biased genes conserved among diatoms and the brown alga *Ectocarpus siliculosus*?

Chapter 2 illustrates the results of a protein based BLAST analysis carried out on the available diatom genomes and selected transcriptomes and in the stramenopile *E. siliculosus,* to detect their conservation degree in other species.

Which is the function of the five MT-biased genes? In which pathways are they involved?

Chapter 3 illustrates the behaviour of the five MT-biased genes during early stages of sexual reproduction analysed on sexualized strains of *P. multistriata.* Their expression trend was higher in sexualised samples against controls and in sexualised samples their expression increased in a time-dependent manner. These results are proof of a MT-specific regulation of the gene expression in sexually competent strains, further supported by the absence of the transcript in samples above the SST. A 24 hours' time course experiment was conducted to test their expression trend and detect possible regulatory mechanisms attributable to the Light:Dark phases and/or the cell cycle.

Is the MT locus conserved among diatoms?

Chapter 4 illustrates a comparative analysis between the planktonic pennate *Pseudo-nitzschia multistriata* and the benthic pennate *Seminavis robusta* focused on HEL-SAM, a gene that was considered part of the MT-locus of *S. robusta.*

Is the mating type genetically determined?

Chapter 4 illustrates the preparation of libraries to be used for running a Bulked Segregant Analysis (BSA). This analysis will be carried out to detect the MT locus in *P.multistriata* and validate its genetic origin.

Chapter 2

The challenge to discover the mating type locus in *Pseudo-nitzschia multistriata*. A transcriptomic approach

2.1 Introduction

RNA-Seq and transcriptomic applications

A transcriptome is the complete set of messenger RNA (mRNA), noncoding RNA (ncRNA) and small RNAs transcripts produced by a particular biological sample (cell, strain, or organism) in a specific condition (Morozova *et al.* 2009). RNA-Seq is a recently developed approach to transcriptome profiling using next generation sequencing technologies. The specific aims of RNA-Seq studies are: i) to detect all transcripts of the biological sample in a specific experimental condition or developmental stage; ii) to determine the transcriptional structure of genes, in terms of their start sites, 5'- and 3'-ends, novel transcribed regions, splicing patterns and other post-transcriptional modifications; and iii) to quantify the expression levels of each transcript (Wang *et al.* 2009). Transcriptomes can thus be very useful to study the molecular machineries used by an organism, cell or strain in defined conditions; they also allow new gene discovery and are helpful for gene annotation of whole genomes.

The process of assembling a transcriptome is challenging, even more if it belongs to a non-model species and it is a *de novo* assembly. To convert raw RNA-Seq data into transcript sequences, one generally aligns reads to a reference genome. However, those methods are unsuitable for organisms with a partial or missing reference genome (Grabherr *et al.* 2011), and for these reason several tools have been developed for the *de novo* assembly of RNA-Seq. Trinity is a method for an efficient and robust *de novo* reconstruction of transcriptomes. It counts three software modules (Inchworm, Chrysalis and Butterfly) applied sequentially to process the RNA-Seq reads (Haas *et al.* 2013).

Once generated a *de novo* RNA-Seq assembly, the following step is the characterization of the transcriptome by annotating all the gene functions. Transcriptomic sequences may be used as an assembly template for further in-depth transcriptome re-sequencing, to develop molecular markers and for gene expression profiling. They can be very convenient in

functional comparisons between different sexes, life stages or tissues within the same organism or different ones (Ekblom and Galindo 2011).

Many physiological questions nowadays find their answers in transcriptomic analyses, for example the organism's response to nutrients starvation (Dyhrman *et al.* 2012, Lauritano *et al.* 2015), different light regimes (Park *et al.* 2010) and thermal stress (Hwang *et al.* 2008). An example of comparative transcriptomic was provided by Di Dato *et al.* (2015) where the main metabolic pathways of three diatom species, *Pseudo-nitzschia arenysensis*, *P. delicatissima* and *P. multistriata* were analysed. A transcriptome could be also used to define the differences between life cycle phases, such as the haploid and diploid stages of the haptophycean *Emiliania huxleyi*, permitting the identification of genes involved in diploid-specific biomineralization, haploid-specific motility, and transcriptional control (von Dassow *et al.* 2009). Another example is the study conducted on the cell cycle phases of the diatom *Seminavis robusta,* related to key cellular processes as chloroplast development (Gillard *et al.* 2008). Moreover a very innovative use of transcriptome analysis finds its application in studying the interaction between a *Pseudo-nitzschia* species and the associated bacteria (Amin *et al.* 2015).

A transcriptomic approach has been used to identify the differences in gene expression between opposite sexes, so identifying a number of sex-biased genes in macro and micro-algae (Martins *et al.* 2013, Patil 2014, Lipinska *et al.* 2015). Examples of sex-biased genes analyses will be reported and discussed in the discussions section of this chapter (Chapter 2.4.1).

When RNA-Seq is used to investigate gene expression changes between alternative conditions, a second independent technique is generally used to validate results. Quantitative real-time polymerase chain reaction (qRT-PCR) is one of the methods mostly used for fast, accurate, sensitive and cost-effective gene expression analysis (Siaut *et al.* 2007). A strict quality control has to be applied throughout the entire procedure as

suggested by many authors (Pfaffl *et al.* 2002, Fleige and Pfaffl 2006, Derveaux *et al.* 2010).

In this chapter, I will focus on the gene expression analysis conducted to differentiate the transcriptomic profile of two mating types (MT+ and MT-) of the marine planktonic diatom *Pseudo-nitzschia multistriata*. The transcriptomes have been sequenced within the project 'A deep transcriptomic and genomic investigation of diatom life cycle regulation' funded by the Joint Genome Institute (http://genome.jgi.doe.gov/Adeeregulation/Adeeregulation.info.html). The project aim was to sequence the transcriptome of two pennate diatoms with similar life cycle features but distinct ecological niches, the planktonic *Pseudo-nitzschia multistriata* and the benthic *Seminavis robusta,* in order to identify genes expressed in different mating types and during distinct phases of the sexual reproduction.

2.2 Material and Methods

2.2.1 Transcriptome samples

The transcriptome was assembled combining RNA-Seq data of four different strains, two MT+ and two MT-, collected in the exponential growth phase. Strains Sy373 and Sy379 were isolated from the Gulf of Naples in 2009 and both strains were collected for RNA extraction when they were below the sexual size threshold (<SST). Strains B856 and B857, belonging to an F2 inbred generation deriving from Sy373 and Sy379 (Fig. 1.6), were collected twice, below the threshold size for sexualisation and above it (Table 2.1).

Table 2.1: Strains of *Pseudo-nitzschia multistriata* used to generate the transcriptome. For each strain are reported: strain code, mating type, size (S= small, L= large) and isolation date.

Strain code	**Mating type (MT)**	**Size**	**Isolation date**
Sy379	MT-	S (43.9 μm)	07/07/2009
Sy373	MT+	S (39.0 μm)	07/07/2009
B857	MT-	S (57.5 μm)	02/08/2011
B856	MT+	S (36.3 μm)	02/08/2011
B857	MT-	L (80.5 μm)	02/08/2011
B856	MT+	L (82.0 μm)	02/08/2011

2.2.2 Sample collection and RNA extraction

Sample collection and RNA extraction were performed in 2012, before the beginning of my PhD. *Pseudo-nitzschia multistriata* cells were grown in f/2 medium (Guillard 1975) at 18° C, an irradiance of 50 mol photons m^{-2} s^{-1} provided by cool-white fluorescent bulbs, and a 12L:12D h photoperiod. Cell growth was monitored by estimating cell concentration

using a Malassez counting chamber. Cells were collected in exponential phase by filtration on 1.2 µm nitrocellulose membranes (Millipore RAWP04700, Billerica, MA, USA). Filters were flash frozen in liquid nitrogen and stored at -80°C. RNA extraction was performed according to TRIzol® protocol (Roche, Basel, Switzerland). Genomic DNA contamination was eliminated digesting with DNase I (Qiagen) according to the manufacturer's instructions followed by RNA clean-up using RNeasy Plant Mini Kit (Qiagen, Venlo, Limburgo, Nederlands). RNA was analyzed by gel electrophoresis (1% agarose w/v) and concentration and quality were determined using a NANODROP (ND 1000 Spectrophotometer), a Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA).

2.2.3 Library preparation and sequencing

Library preparation and sequencing were done at the JGI in 2012 before the beginning of my PhD. Poly-A RNA was isolated from 5µg total RNA using Dynabeads mRNA isolation kit (Invitrogen - Life Technologies, Carlsbad, CA, USA). Isolation procedure was repeated to ensure that the sample was depleted of rRNA. Purified RNA was then fragmented using RNA Fragmentation Reagents (Ambion- Life Technologies, Carlsbad, CA, USA) at 70 °C for 3 mins, targeting fragments range 200-300 bp. Fragmented RNA was then purified using Ampure XP beads (Agencourt - Beckman Coulter Genomics, Danvers, MA, USA). Reverse transcription was performed using SuperScript II Reverse Transcription (Invitrogen - Life Technologies, Carlsbad, CA, USA) with an initial annealing of random hexamer (Fermentas - Life Technologies, Carlsbad, CA, USA) at 65 °C for 5 mins, followed by an incubation of 42 °C for 50 mins and an inactivation step at 70 °C for 10 mins. cDNA was then purified with Ampure XP beads. This was followed by second strand synthesis using dNTP mix, where dTTP is replaced by dUTP. Reaction was performed at 16 °C for 1 h. Double stranded cDNA fragments were purified and selected

for targeted fragments (200-300 bp) using Ampure XP beads. The dscDNA were then blunt-ended, poly- adenylated, and ligated with library adaptors using Kapa Library Amplification Kit (Kapa Biosystems Inc, Wilmington, MA, USA). Adaptor-ligated DNA was purified using Ampure XP beads. Digestion of dUTP was then performed using AmpErase UNG (Applied Biosystems - Life Technologies, Carlsbad, CA, USA) to remove second strand cDNA. Digested cDNA was again cleaned with Ampure XP beads. This was followed by amplification by 10 cycles PCR using Kapa Library Amplification Kit (Kapa Biosystems Inc, Wilmington, MA, USA). The final library was cleaned with Ampure XP beads. Sequencing was done on the Illumina HighSeq platform generating paired end reads of 150bp each.

2.2.4 Sequencing data analysis

The bioinformatic analyses to assemble and annotate the transcriptome and to identify the differentially expressed transcripts were performed by Dr. Remo Sanges (SZN).

Transcriptome assembly

Raw reads were filtered and trimmed based on quality and adapter inclusion using Trimmomatic (Lohse *et al.* 2012) with the following parameters: -threads 20 -phred64 ILLUMINACLIP:illumina_adapters.fa:2:40:15 LEADING:5 TRAILING:5 SLIDINGWINDOW:5:20 MINLEN:100. Trimmed and filtered reads were normalized using the normalize_by_kmer_coverage.pl script from the Trinity (Grabherr *et al.* 2011) software release r2013_08_14 with the following parameters: --seqType fq --JM 220G --max_cov 30 --SS_lib_type RF --JELLY_CPU 22. The assembly was performed using Trinity on the trimmed, filtered and normalized reads with the following parameters: --

seqType fq --JM 220G --inchworm_cpu 22 --bflyHeapSpaceInit 22G --bflyHeapSpaceMax 220G --bflyCalculateCPU --CPU 22 --SS_lib_type RF --jaccard_clip --min_kmer_cov 2.

Annotation

To annotate the translated transcriptome, BLASTX and RPSBLAST searches were run to align sequences against proteins and domains (Camacho *et al.* 2009). BLASTX was used to align putative protein sequences against the sequences in Swiss-prot (SP) and Uniref90 (Suzek *et al.* 2007). The databases have been downloaded in August 2013, the number of sequences in Uniref90 was 15,996,810 and 540,958 in Swiss-prot. The parameters chosen for the BLASTX were: word_size = 4 evalue = 10−5 num_descriptions = 5 num_alignments = 5 threshold = 18. For each sequence the best hit, if any, was chosen and associated to the transcript. RPSBLAST search was used to annotate the domains composition of the transcripts against the Conserved Domains Database (CDD) collecting multiple sequence alignment models for domains from Pfam, SMART, COG, PRK, TIGRFAM (Marchler-Bauer *et al.* 2011). Search parameters used: evalue = 10−5 num_descriptions = 20 num_alignments = 20. All the collected results (domain id, name, start and end, e-value and description) are parsed from the RPSBLAST output and added to the final table. Mapping of Gene Ontology (GO) functional classification (Ashburner *et al.* 2000) and the Enzyme Commission IDs and descriptions (Bairoch 2000) were performed mapping the SwissProt ID of the best matches in the tables idmapping_selected.tab from the Uniprot distribution and enzyme.dat from the Expasy database. Annocript was used to achieve the graphical output of the most enriched family groups annotated (Musacchia *et al.* 2015).

Reads mapping

Raw reads were mapped on the transcriptome using bowtie (Langmead *et al.* 2009) with the following parameters: -p 24 --chunkmbs 10240 --maxins 500 --trim5 20 --trim3 20 --seedlen 20 --tryhard -a --nofw. Sam output file from bowtie were converted in bam, sorted, indexed and counted using respectively the sort, index and idxstats programs from the samtools collection (Li *et al.*, 2009). We have generated a final table containing the number of reads mapping on each transcript from each sample using a custom R script on the output of the samtools idxstats program. All the transcripts showing less than 0.5 reads mapping per million mapped reads (cpm) in more than 4 samples were discarded from the transcriptome as being too lowly expressed and hence probably deriving by transcriptional noise or procedural artefacts.

Differential expression analysis

Counts were loaded into the R environment and the Bioconductor edge R package (Robinson *et al.* 2010) was used in order to select transcripts significantly differentially expressed between the two mating types. The analysis performed the following steps: calculation of the normalization factors using the calcNormFactors function, estimation of the common dispersion using the estimateCommonDisp function, estimation of tagwise dispersion using the estimateTagwiseDisp function, statistical test for differential expression using the exactTest topTags functions with p-value FDR correction. We considered as significantly differentially expressed all those transcripts showing a linear fold change of +/-2 and an FDR corrected p-value smaller or equal to 0.1.

2.2.5 Transcripts identification and BLAST analysis

From here onwards I have carried out the work.

Expression levels of differentially expressed transcripts were compared and a sub-set of putative MT-biased transcripts was selected according to different criteria:

1. divergent expression rates between MT+ and MT- transcripts;
2. statistical significance of the differential expression (corrected p-values);
3. annotated function of the protein encoded by the transcripts;
4. the shared presence in both *Seminavis robusta* and *Pseudo-nitzschia multistriata*.

The selected transcripts were checked with a BLASTN on the *P. multistriata* genome to verify the correspondent gene model and define the exon/intron structure, then with TBLASTX/BLASTP searches in public databases to verify reliability of the annotation.

2.2.6 Primer design

Primers for the selected transcripts were designed manually using EditSeq software (DNASTAR Inc.). The following criteria were used to design the primers: the GC content as close as possible to 50%, primer length between 19 and 22 bp and amplicon size between 100 and 200 bp. Tm calculator (http://www6.appliedbiosystems.com/support/techtools/calc/) was used to calculate the optimal melting temperatures of primers. Specificity of the primers was checked through a BLASTN on the *P. multistriata* genome and transcriptome. A total of 63 primers was designed and tested for all the selected candidates. In Table 2.2 are listed the primer pairs used for quantitative real time PCR validation.

Table 2. 2: List of primers for the transcripts validated through qRT-PCR. The transcript name, primer name, primer sequences and amplicon size are reported.

Transcript ID	**Primer name**	**Sequence**	**Amplicon size (bp)**
Locus11310v1rpkm7.58	Pm7.58F Pm7.58R	5'-GCAGCTGAAATTCACGTAG-3' 5'-GGTGTTGTTTGTAGCGTTATC-3'	153
Locus1124v1rpkm204.78	Pm204.78F	5'CTGAAGATGCTTGCTGTTCG-3'	147

	Pm204.78R	5'-TCCAACGACTCTTGCACTTG-3'	
Locus1771v1rpkm127.15	127.15 F3 127.15 R3	5'-CCTCCGAATATGGATACATG-3' 5'-GAGCTAAACATCGTGACACC-3'	194
Locus11029v1rpkm8.14	Pm8.14F Pm8.14R	5'-GGGAGAGTGAAGAATGTGGTTAG-3' 5'-CATTCTGCTTGTTTTGTGACG-3'	146
Locus27553v1rpkm6.74	6.74+F 6.74+R	5'-GTACCTGACAGCACTCATACCG-3' 5'-GCACTACACTTTCTGTTTCGGTC-3'	115
Locus26972v1rpkm7.30	7.30+F 7.30+R	5'-CCGTCGAAGTCTTCGTTG-3' 5'-CGACCTCGGTACTTACGC-3'	177
Locus25079v1rpkm9.31	9.31+F 9.31+R	5'-CCGTTGCAAACCTGATCG-3' 5'-GATATCCAGCCGAGGATGC-3'	156
Locus52839v1rpkm0.00	0.00+F 0.00+R	5'-GTATGGCGCTCACCACTTC-3' 5'-CGTCTTCGACTGCGTCTTC-3'	156
Locus21788v1rpkm13.42	13.42-F 13.42-R	5'-GGAGCTCTATCCGATCGAGTC-3' 5'-CAACTGCGCATCGATGATTC-3'	204
Locus20443v1rpkm15.38	15.38-F 15.38-R	5'-GCGGTCCAACACTAACGATC-3' 5'-CGATGAAACCACAAAGTTTCG-3'	147
comp55263_c0_seq12	FAS F FAS R	5'-GATGACACCGTCGCAAGC-3' 5'-GGTTGTGGCGAGTCCCTC-3'	159
comp55637_c0_seq16	RAS F RAS R	5'-GCCGACGGAATCATCATG-3' 5'-CGGACTTGTTTCCGACCAG-3'	130
comp55637_c0_seq16	55637F 55637R	5'-GTCGCGGTTAGAAATCGTC-3' 5'-GTTCGCTAGGAATAGTGCCC-3'	115
comp55333_c3_seq3	HMGB F HMGB R	5'-CTTCCCCCAAAGGCACTG-3' 5'-CAAAGCCAGTCGCTGTCATC-3'	134
comp55333_c3_seq3	bHLH F bHLH R	5'-GCAGTCTCCGGACAACCTC-3' 5'-GACTCCACTCGTCTCACCCTC-3'	172
comp53977_c1_seq12	CAP_ED F CAP_ED R	5'-GGAAAGGTCGAGGTCCAGAC-3' 5'-GATCTCGACCACGTCCACC-3'	147
comp53977_c1_seq2	CAP_ED3 F CAP_ED3 R	5'-GCTCGTTCTTGGTGTCGATG-3' 5'-GGAGACTTTTTCGGGGAGG-3'	148
comp47507_c2_seq2	47507F 47507R	5'-CCCCTACAAGCTCTTTGATTTG-3' 5'-GAAATTGTGGTGCCCAAAG-3'	160
comp43946_c0_seq6	43946F 43946R	5'-GTGGTGCGGGCACTGCAAGG-3' 5'-GTCGGTACAGTCGACGCTTC-3'	103
comp46228_c0_seq3	46228F 46228R	5'-CCACCGAACTAGGCAACTGTC-3' 5'-GGCACAGAACCCGTCAAC-3'	139
comp22480_c0_seq3	22480F 22480R	5'-GCCGCAGCTTATTGACTGAAC-3' 5'-CCTTTCTTTGGGAGTTGAAAGC-3'	183
comp6261_c0_seq1	6261F 6261R	5'-CGACAACATCATCCTTCCAC-3' 5'-GTGCCACCGACTGTAACAAG-3'	163
Locus26783v1rpkm7.51	7.51-F	5'-CAGGAACAGCCGATGCTTC-3'	187

	7.51-R	5'-CAAGAATGGCCGAGACGAC-3'	
comp27491_c0_seq3.1	27491- F 27491- R	5'- GCAAAAGGCAAAGAAGATCAT-3' 5'- GAGGGTGTGTGGAGATAGTTTG-3'	156
comp27022_c0_seq1.1	0081900- F 0081900- R	5'-CCGATCACCAGTATCCCATC-3' 5'-CAATAGGCGCCCTAGGAAT-3'	150
comp25269_c0_seq1	0093550.1- F 0093550.1- R	5'-GAAAAATGAATCCCGACACC-3' 5'-GTATTGCTGGTTTTGGTTCG-3'	143
comp29120_c0_seq2.1	29120- F 29120- R	5'-CGAGCACGAGTATGATTCTGTAG-3' 5'-CTTGTTGGCACGCCTCTAT-3'	170
comp20279_c0_seq4.1	0020770+ F 0020770+ R	5'-GCGCAAGCAATCTAAGGTG-3' 5'-GACGTCGACGGCTATTTTG-3'	166
comp31481_c0_seq1.1	0121970+ F 0121970+ R	5'-GATTCCGATTTCGAGAAACC-3' 5'-CGATTTGGAGTCGAGAAAGG-3'	177

2.2.7 Cultures

The strains of *Pseudo-nitzschia multistriata* used for the experimental work of PCR validations were established from the isolation of single cells from net samples collected at the LTER-MC station in the Gulf of Naples between 2010 and 2013 or obtained by crosses carried out in the laboratory (F1 = Sy776-*SP2+) (Table 2.3). Two sets of unrelated strains were used; the first one was used to perform a first series of PCR validations during the first part of the study, the second was used for the final qPCR validation, culture treatment and RNA extraction was carried out at the same time for all the strains, unrelated to those used for the RNA-seq experiment

Table 2.3: Strains of *Pseudo-nitzschia multistriata* used for the PCR validations experiments. For each strain are reported: the strain code, the mating type, the isolation date and the RNA extraction date.

Strain code	**Mating type (MT)**	**Isolation date**	**RNA extraction date**
First set of validations			
B857	MT-	02/08/2011	04/05/2012
A8	MT-	13/10/2010	20/09/2012
A3	MT-	13/10/2010	16/04/2012

Sy680	MT-	07/09/2010	11/01/2013
Sy686	MT-	07/09/2010	11/01/2013
Sy679	MT-	07/09/2010	11/01/2013
B856	MT+	02/08/2011	04/05/2012
B854	MT+	13/10/2010	16/04/2012
Sy793	MT+	21/09/2010	11/01/2013
Sy710	MT+	07/09/2010	11/01/2013
Sy673	MT+	07/09/2010	20/09/2012
Second set of validations			
B935	MT+	24/05/2012	13/12/2013
SH18	MT+	07/04/2013	13/12/2013
MVR1041.6	MT+	05/02/2013	13/12/2013
MVR171.8	MT+	07/06/2013	13/12/2013
B936	MT-	24/05/2012	13/12/2013
SH20	MT-	07/04/2013	13/12/2013
MVR1041.4	MT-	05/02/2013	13/12/2013
MVR171.1	MT-	07/06/2013	13/12/2013

The cultures were grown in f/2 culture medium (Guillard 1975) prepared with oligotrophic seawater collected offshore in the Gulf of Naples. The sea water was filtered over a 0.45 m pore-size nitrocellulose membrane filter (Millipore S.p.A., Milano, Italy) and then autoclaved. Salinity was adjusted to 36 by adding sterile milli-Q water, and f/2 was obtained through addition of 20 ml of 50x concentrated f/2 (Sigma Aldrich., St. Louis, MO, USA) per litre. The f/2 medium was filtered over a 0.22 m pore-size Filter Stericup GP SCGPU05RE (Millipore, Billerica, MA, USA) just before the use, in order to eliminate precipitates.

Strains were maintained in a growth chamber at a temperature of 18 °C, a photoperiod of 12:12 h Light:Dark, and a photon flux density of 50-60 μmol photons $m^{-2}s^{-1}$ provided by cool white fluorescent tubes TLD 36W/950 (Philips, Amsterdam, Nederland).

2.2.8 Mating experiments

The only way to infer their mating type is to isolate single cells, establish clonal cultures and carry out a matrix of crosses. The results will allow the identification of strains of opposite mating type, since an MT+ will only produce sexual stages when crossed with MT- and *viceversa*. However, we cannot state which strain is the MT+ and which one is the MT-. Nevertheless, we can rely on the conjugation modality of this model species to establish a conventional system that allows having a consistent criterion to attribute mating type. The gametes produced by one gametangium are in fact active and conjugate with those produced by the passive gametangium that bears the zygotes and subsequently the auxospores. If crosses are carried out using strains with different length, it is possible to differentiate the mating type and conventionally attribute the MT- to the strain that carries the auxospores and the MT+ to the strain that produces the 'migrating' gametes. In this way reference strains of opposite mating types can be identified and used to carry out crosses with larger number of strains.

To assess or to confirm the mating type of the strains used for the qRT-PCR analyses, strains were crossed pairwise following the protocol described above. The matrix of crosses included one pair of reference strains of known mating type. A few drops of exponentially growing culture were inoculated in 6-wells Costar tissue culture plates (Corning Inc., New York State, USA) filled with 5 ml of f/2 medium. The cross was incubated in a culture room at a temperature of 21 °C, a photon flux density of 100-130 µmol photons $m^{-2} \cdot s^{-1}$ provided by cool white fluorescent tubes and natural light coming from a window facing North and a natural photoperiod. The culture plates were inspected every day with a Leica DMIL inverted microscope (Leica Microsystems, Wetzlar, Germany) to check for the presence of zygotes and auxospores.

2.2.9 Sampling, RNA extraction and reverse transcription

Pseudo-nitzschia multistriata cells were grown as illustrated in the 2.2.7 'Cultures' section. Cell growth was monitored by estimating cell concentration using a Sedgewick-Rafter counting chamber and cells were collected when in exponential growth phase (~ 100,000 cell·ml^{-1}) by filtration on 1.2 µm pore size nitrocellulose membranes RAWP04700 (Millipore, Billerica, MA, USA). Filters were submerged in 1.5 ml TRIzol®, flash frozen in liquid nitrogen and stored at -80 °C.

RNA extraction was performed according to TRIzol® protocol (Roche, Basel, Switzerland). Genomic DNA contamination was eliminated digesting with DNase I (QIAGEN) according to the manufacturer's instructions followed by RNA clean-up using RNeasy Plant Mini Kit (Qiagen, Venlo, Limburgo, Nederlands). RNA was analysed by gel electrophoresis (1% agarose w/v), a Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) to assess concentration and a NANODROP (ND 1000) spectrophotometer to determine the quality as 260/280 nm and 260/230 nm absorbance ratios. RNA contamination by genomic DNA was tested with PCR amplification.

One µg of the total RNA extracted was used for cDNA preparation using the QuantiTect® Reverse Transcription Kit (Qiagen, Venlo, Limburgo, Nederlands). cDNA integrity was assessed amplifying a 1kb fragment containing an intron of the reference gene TUB A with primers TUB A Fw intron: 5'-CGAGAGTAACCTTTAAATGCCAAG-3' and TUB A Pm rv: TUB A Pm rv 5'-GACGACATCTCCACGGTAC-3'.

2.2.10 PCR and quantitative real-time PCR validations

PCR experiments were conducted on both genomic DNA and cDNA for MT+ and MT- samples. The reactions were done in final volumes of 20 µL: cDNA 1 µL, oligo fw (2.5 µM), oligo rv (2.5 µM), PCR reaction buffer with $MgCl_2$ 10X (Roche, Basel, Switzerland),

dNTP (2 mM), Taq DNA Polymerase (0.25 U/μL) (Roche, Basel, Switzerland). The thermal profile of amplification varied depending on the fragment to be amplified. The products were checked on 1.5 % agarose gel with a 100 bp marker to recognize the size of the band amplified (Ladder 100 plus. Fermentas - Life Technologies, Carlsbad, CA, USA). RT-PCR analyses were performed to confirm expression and to operate a first sorting on the transcripts showing a real difference between mating types. The transcripts that stood the test were then validated through qRT-PCR.

Two sets of qRT-PCR validations

PCR analyses were performed twice on two sets of independent strains of *P. multistriata* within the size for sexualisation, both containing MT + and MT - (Table 2.3). The first set was composed by 6 MT- and 5 MT+ samples. RNA extractions had been performed between 2012 and 2013 by different operators. To obtain a set of samples that were processed all at the same time, in order to minimize variability, I produced RNAs from four additional MT- and four MT+ samples. I collected the RNA from eight strains of *P. multistriata,* when in exponential growth phase and at similar cell concentration (Table 2.3). Cells were concentrated the same day and at the same hour of the day (± 1 hr). RNA quality was checked and considered suitable for qRT-PCR analyses. Quality controls for the cDNA were performed as described in section 2.2.9.

qRT-PCR conditions

Real time qPCR amplification was performed using 1 μL of a diluted cDNA, 4 μL of the primers (final concentration 0.7 μM of each primer) and 5 μL of Fast SYBR Green Master mix with ROX (Applied Biosystems by Life Technologies, Carlsbad, CA, USA) in a final volume of 10 μL, using ViiA™ 7 Real-Time PCR System (Applied Biosystems by Life Technologies, Carlsbad, CA, USA). Each sample was analyzed in technical triplicate to

capture intra-assay variability and each assay included at least two negative controls for each primer pair. PCR conditions were as follows: 95 °C for 20 s, 40 cycles at 95 °C for 1 s and 60 °C for 20 s, 95 °C for 15 s, 60 °C 1 min, and a gradient from 60 °C to 95 °C for 15 min.

Primers specificity and efficiency

Primer amplification efficiency was calculated with a serial 10-fold dilution using Standard Curve method of ViiA™ 7 Real-Time PCR System (Applied Biosystems by Life Technologies, Carlsbad, CA, USA). The results were double checked on Excel manually, calculating the slope and efficiency of each primer pair. The calibrator's cDNA sample used was obtained by mixing together cDNA of eight independent samples (4 MT+ strains and 4MT- strains). Five cDNA dilution points have been chosen: 1:1, 1:5, 1:10, 1:50 and 1:100. The PCR conditions used were those reported in the previous paragraph. Primer pairs used for the experiment had efficiency comprised between $1.75 > E < 2.1$. In addition, the specificity of the PCR products was verified by melting-curve analysis for all transcripts tested, discarding the ones with double peaks and evident primer-dimers.

Reference genes

The analysis of the reference genes for *P. multistriata* was conducted by Adelfi *et al.* (2014). Nine housekeeping genes were tested genes for stable expression under different experimental conditions. From a NormFinder and geNorm-based analysis, it was found that only *TUB A, TUB B* and *CDK A* were genes stable in all the conditions analyzed. However, in addition to these three genes, also *ACT* and *COPA* were included in the group of good reference genes for *P. multistriata* (Adelfi *et al.* 2014). The reference genes I selected for the qRT-PCR validations were *TUB A, TUB B* and *CDK A.* They showed to be the less variable, i.e. they were not changing among MT+ and MT- samples. Previous to

this decision, I tested also *H4, TBP* and *RPS* as reference genes of the first cDNA set of samples for qRT-PCR validations. *TBP* resulted to be adequate for the validations but still *TUB A, TUB B* and *CDK A* were the best choice.

REST - qPCR data analysis

Expression analysis was performed using the Relative Expression Software Tool-Multiple Condition Solver (REST-MCS), the calculation software for the relative expression in qRT-PCR, using Pair Wise Fixed Reallocation Randomization Test (Pfaffl *et al.* 2002). The relative expression ratio was calculated from the real-time PCR efficiencies and the crossing point deviation of an unknown sample versus a control (CP value) (eq. 1) (Pfaffl *et al.* 2002). The relative expression ratio (R) of the targeted mating-type related genes was computed as the expression variation between one mating-type, set as control, against the other mating-type, set as condition, normalized over the expression variation of reference genes whose expression levels were not regulated in specific experimental conditions.

(eq. 1) $$\text{Ratio (R)} = (E_{target})^{\Delta CP\ target\ (control-sample)} / (E_{ref})^{\Delta CP\ ref\ (control-sample)}$$

Equation 1: Is the equation employed by REST to calculate the relative expression variation of a target gene, where: **E** is the specific efficiency calculated for each gene, **CP** is the Crossing Point for each gene in the different conditions, $\mathbf{E_{target}}$ is the real-time PCR efficiency of target gene transcript, $\mathbf{E_{ref}}$ is the real-time PCR efficiency of a reference gene transcript, $\mathbf{\Delta CP_{target}}$ is the CP deviation of control − sample of the target gene transcript, and $\mathbf{\Delta CP_{ref}}$ is the CP deviation of control − sample of reference gene transcript.

PCR and sequencing to test MRM1 duplication

To analyse the flanking regions of gene *MRM1*, where an in/del was observed, two primers were designed upstream the start site of the gene in the putative promoter region to amplify and sequence a fragment of 728 bp. The primers were designed manually using EditSeq software (DNASTAR Inc.) (Sc432promFw: GAGTTCTCTTGCCGGATGATAC; Sc432promRv: CCCTCATTCACCACATGTGAC).

The PCR experiments were conducted on genomic DNA of seven *P. multistriata* strains (Table 2.4).

Table 2.4: Strains of *P. multistriata* used to sequence the promoter region of *MRM1*. Reported in the table are the strain code and mating type.

Strains ID	MTs
1078-30	MT+
1120-47	MT-
1075-25	MT+
1119-15	MT+
1120-32	MT-
1120-48	MT-
VF2.2	MT-

PCR reactions were carried out in a volume of 100 µl: gDNA 2,5 µL, oligo fw (2.5 µM), oligo rv (2.5 µM), PCR reaction buffer with $MgCl_2$ 10X (Roche, Basel, Switzerland), dNTP (2 mM), Taq DNA Polymerase (0.25 U/µL) (Roche, Basel, Switzerland). The thermal profile of amplification varied depending on the fragment to be amplified. The products were checked on 1 % agarose gel in TAE buffer and ethidium bromide staining with a 1 Kb ladder, to recognize the size of the band amplified (Gene Ruler 1 kb DNA Ladder - Thermo Scientific Fermentas, Waltham, Massachusetts, USA). The PCR products were purified with QIAquick PCR purification kit (Qiagen, Venlo, Limburgo, Nederlands) according to the manufacturer's instructions. The sample for the sequencing reaction was composed by purified DNA [15 fmol/µl] + primer [4,5 pmol/µl] in a final volume of 20 µl. Sequence reactions were obtained with the BigDye Terminator Cycle Sequencing technology (Applied Biosystems, Foster City, CA), purified in automation using the Agencourt CleanSEQ Dye terminator removal Kit (Agencourt Bioscience Corporation, 500 Cummins Center, Suite 2450, Beverly MA 01915 - USA) and a robotic station Biomek FX (Beckman Coulter, Fullerton, CA). Products were analyzed on an Automated Capillary Electrophoresis Sequencer 3730 DNA Analyzer (Applied Biosystems).

Editing and alignment of the sequences was conducted with SeqMan software (DNASTAR Inc.).

2.2.11 BLAST analyses

The transcripts resulted differentially expressed between the MT+ and MT- of *P. multistriata* were deeply analyzed studying the nucleotidic and proteic sequence.

The nucleotidic sequences were aligned against the reference genome of *P. multistriata* http://gbrowse255.tgac.ac.uk/cgi-bin/gb2/gbrowse/maplesod_psnmu_v1_4_gbrowse255/ to study the gene structure. Then the nucleotidic sequences were translated to protein sequences with the ExPASy translate tool http://web.expasy.org/translate/, identifying the correct open reading frames (ORF) among the six frame translations. EditSeq software (DNASTAR Inc.) was used to predict the molecular weight of the protein, expressed in Daltons. To confirm the functional annotation of the transcripts produced during the transcriptome annotation (Chapter 2.2.4), conserved domains were searched through http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. The protein was also checked with SMART sequence domain identifier tool (Simple Modular Architecture Research Tool) http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1 (Letunic *et al.* 2015), which allows the identification and annotation of genetically mobile domains and the analysis of domain architectures thanks to an underlying non-redundant protein database synchronized with UniProt, Ensembl and STRING. Moreover HMMER searches of the SMART database occur by default while searches for outlier homologues and homologues of known structure, PFAM domains, signal peptides, internal repeats and intrinsic protein disorder (protein that lacks a fixed or ordered three-dimensional structure) can be selected in the set up preferences. SignalP 4.1 was used to discriminate signal peptides from transmembrane regions (Petersen *et al.* 2011) and ASAFind version 1.1.6. to detect eventual nuclear-encoded plastid-localized proteins (Gruber *et al.*, 2015).

WoLF PSORT http://www.genscript.com/psort/wolf_psort.html was used to search for protein localization (Horton *et al.* 2007). This program is an extension of the PSORT II program for protein subcellular localization prediction based on sorting signals, amino acid composition and functional motifs. However the WoLF PSORT dataset is divided into fungi, plant and animal so it is not specifically optimized for diatoms. Thus, in this case, an approximation of the results achieved from each of the three datasets was made to get an indicative prediction of protein localization.

To search for conservation, the protein of the five sex (MT)-biased genes were blasted through a TBLASTN against the available genomes of 12 heterokonta species among which four diatom species: *Pseudo-nitzschia multiseries* CLN-47, *Fragilariopsis cylindrus* CCMP 1102, *Thalassiosira pseudonana* CCMP 1335, *Phaeodactylum tricornutum* (JGI genomes, http://genome.jgi.doe.gov/), and the genome of the stramenopile macroalga *Ectocarpus siliculosus* (http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi). The five proteins were also blasted against all the transcriptomes produced within the Marine Microbial Eukaryote Transcriptome Sequencing Project (http://marinemicroeukaryotes.org/) and on *Seminavis robusta* within the JGI founded project "A deep transcriptomic and genomic investigation of diatom life cycle regulation" (http://genome.jgi.doe.gov/Adeeregulation/Adeeregulation.info.html). The custom BLAST tool for the MOORE and JGI affiliated species was created by Dr. Remo Sanges on the SZN server. Only the genes corresponding to proteins conserved not only in the functional domain but also in other regions were considered orthologues. The protein products of the orthologous genes were downloaded from their reference genomes or transcriptomes, manually checked for their correct frame translation and validated with a reciprocal TBLASTN on the *P. multistriata* genome and transcriptome browser. The protein sequences were aligned according to ClastalW interface and the multiple alignments were

manually curated with BioEdit v7.2.5 (Tom Hall, Ibis Biosciences, An Abbott company, 2251 Faraday Avenue, Carlsbad, CA 92008).

MEGA6 (Molecular Evolutionary Genetic Analysis software) (Thompson *et al.* 1994) was used on the multiple alignments of the protein sequences to check the degree of protein conservation and conserved motif among the homologs genes.

2.2.12 Ka/Ks calculation

The genome-wide study of selective pressure acting on protein coding genes of *P. multistriata* by means of Ka/Ks (number of non synonymous mutations/number of synonymous mutations) calculation was performed in collaboration with the company Sequentia (http://www.sequentiabiotech.com/). The analysed data included 12,152 and 19,703 CDS sequences of *P. multistriata* and of *P. multiseries* (Psemu1, downloaded from the JGI), respectively.

As a first step, a reciprocal best BLAST hit (RBH) approach was used to identify *P. multistriata* and *P. multiseries* orthologous sequences. Only alignments covering at least 30% of *P. multistriata* sequences were retained. The RBH was calculated using both the e-value and the bit-score of the alignment. The RBH analysis resulted in7,128 reciprocal best BLAST hits between *P. multistriata* and *P. multiseries*.

As a following step each pair of sequences of *P. multistriata* and *P. multiseries* were aligned with Prank (v.150803, http://wasabiapp.org/software/prank/) using empirical codon model and the alignments were refined by using trimAL (v1.4.rev15, http://trimal.cgenomics.org/) to remove gaps and badly aligned regions. Of the 7,128 processed alignments, 6,066 (85%) were suitable for Ka/Ks calculation. Ka/Ks calculation was performed with KaKs_Calculator (v.2, https://sourceforge.net/projects/kakscalculator2/), the model for the calculation was chosen for each alignment by using the AICc model selection method.

2.3 Results

2.3.1 *Pseudo-nitzschia multistriata* transcriptome

The sequencing of mRNA-enriched total RNA, after adapter trimming and quality assessment, yielded between 85 million and 105 million reads for MT+ and between 61 million and 131 million reads for MT- samples. An overview of the seven libraries generated through Illumina sequencing is presented in Table 2.5. All the libraries were used to generate the transcriptome assembly.

Table 2.5: Information on the RNA libraries used for the transcriptome assembly of *P. multistriata*. (*) The two libraries have been subsequently merged.

Library_Name	Sample_Name	Mating type	Size	RawReads Number	Seq platform
CIIO (*)	Sy373	MT+	S	52,847,496	Illumina HighSeq
CIIO (*)	Sy373	MT+	S	85,472,786	Illumina HighSeq
CIIP	Sy379	MT-	S	61,712,734	Illumina HighSeq
HCUH	B856	MT+	S	109,090,252	Illumina HighSeq
HCUN	B857	MT-	S	105,703,908	Illumina HighSeq
HCUO	B856	MT+	L	102,353,212	Illumina HighSeq
HATT	B857	MT-	L	131,284,002	Illumina HighSeq

From now on, I will refer to the six libraries as: S1+→CIIO (the two CIIO libraries were merged), S1-→CIIP, S2+→HCUH, S2-→HCUN, L2+→HCUO, L2-→ HATT.

An overview of the results of the assembly is presented in Table 2.6.

Table 2.6: *P. multistriata* transcriptome. Summary of the general statistics conducted on the assembly.

General statistics	
Total number of transcripts	30691
N50	2109
Average length	1545
Median length	1312
Minimum length	201
Maximum length	20729
Average GC	0.4975
Sequences with at least 1 annotation	20825

The 67% of the sequences had at least one annotation. Functional classification of the most expanded protein groups according to the Pfam annotation is shown in Figure 2.1. The largest protein families in diatoms family is represented by the Pfam00069 Protein kinase domain group.

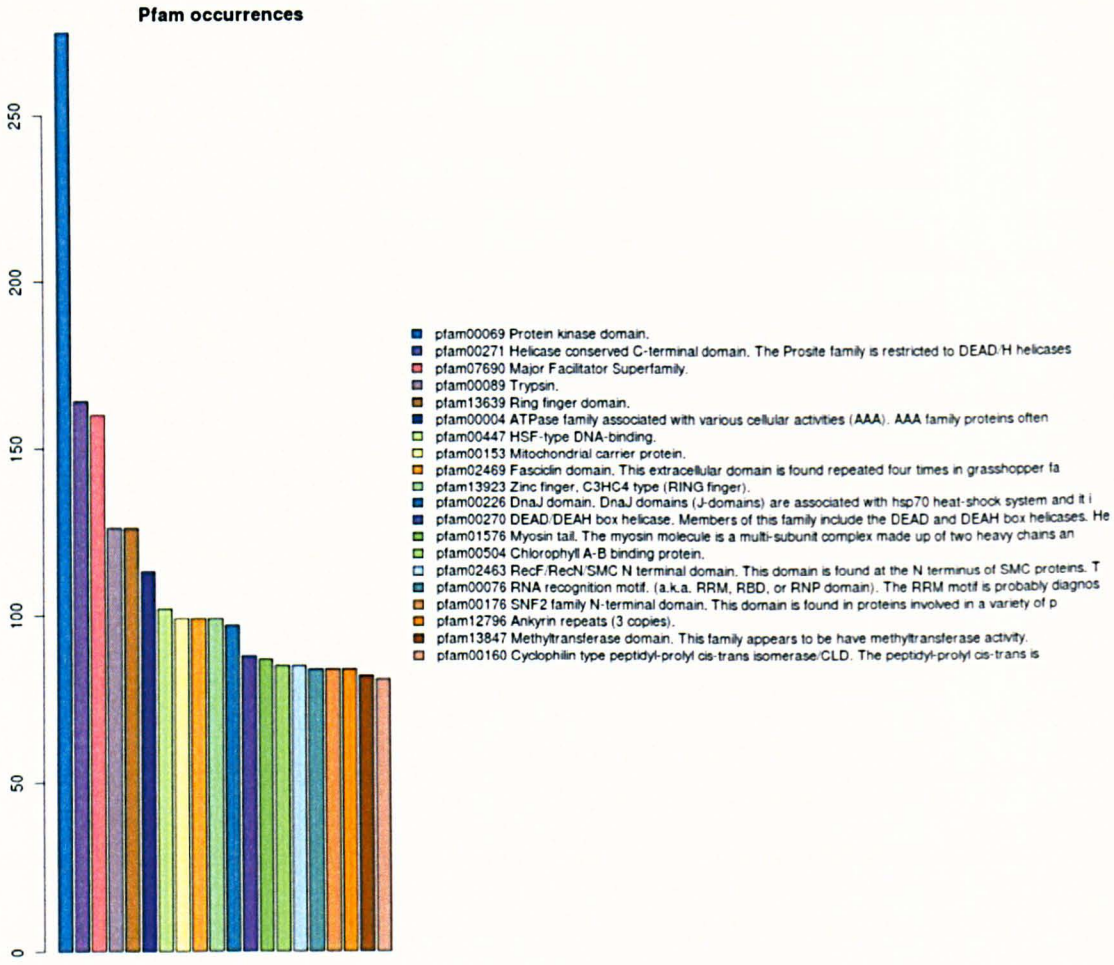


Figure 2.1: Histogram of the occurrence of the annotated transcripts of *P. multistriata* encoding for proteins belonging to major protein families.

2.3.2 Differential expression analysis

The *P. multistriata* transcriptome has been the first attempt of *de novo* assembly with Trinity on a unicellular non model organism. To provide the best *de novo* assembly of the transcriptome, Dr. Sanges performed four bioinformatic analyses, each time improving the algorithm of the software. Four assemblies were produced over time (Table 2.6) and the transcripts have been identified by different codes: the first two (1st and 2nd) have the JGI ID "Locus..." while the second two (3rd and 4th) have as ID "comp...".

A differential expression analysis was performed on each assembly, using the *de novo* transcriptome assembled as reference, to identify those transcripts significantly changing between mating types, considering that the number of raw reads mapped is directly proportional to the mRNA abundance in the tested samples. The differential expression analyses resulted in a list of significantly differentially expressed transcripts between mating types (Table 2.7). The list of all the selected transcripts from the previous assemblies is reported in Table 2.8.

Table 2. 7: The four assemblies of *Pseudo-nitzschia multistriata* transcriptomes and number of differentially expressed genes (DEG).

	Assembly ID	Number of DEG
1st	First assembly	298
2nd	Second assembly	522
3rd	Third assembly	211
4th	Fourth assembly	91

To choose which transcripts were good candidates to be tested as mating type related, I compared the expression level of MT+ and MT- transcripts and selected a set of putative MT-biased transcripts according to the following criteria:

1. Transcripts that highly differed between the MT+ and MT- samples at the level of normalized counts. They were highly expressed in one mating type and completely or partially absent in the other.

2. Low p-values of the expression fold change for the selected transcripts were representative of high statistical significance of the resulted counts.

3. The annotated function of the highest differentially expressed transcripts presented a possible involvement in the system of sex determination and/or signalling, gamete recognition and conjugation. Among the differentially expressed transcripts, I found enrichment for genes involved in DNA rearrangement, cellular adhesion, and membrane trafficking and signalling; those are all processes known to be linked with sex determination system in other organisms. However, about 65% of the list of transcripts had unknown function

I have searched for homologues of a series of loci putatively involved in mating type determination in *Seminavis robusta*, because a sister project on the transcriptome analysis of this benthic pennate diatom was carried out by the research team of prof. W. Vyverman at the University of Ghent (Belgium). The search for homologs between the two species was performed by TBLASTX searches using the *P. multistriata* transcripts as query against the *S. robusta* transcripts. No homology was found.

2.3.3 Old transcriptome assemblies and validations

From the results of the differential expression analyses of the various assemblies, candidate MT-biased genes to be validated were selected following the criteria illustrated above.

The 47 transcripts - selected from the previous assemblies and from the final one, that satisfied the selection criteria, are listed in Table 2.8. The results of the *in silico* differential expression analysis and of the real time quantitative PCR validations are reported in the table and, for each transcript belonging to the earlier analyses, the correspondence with the final assembly has been listed.

Table 2.8: The list of the 47 selected transcripts for each assembly (1st, 2nd and 3rd assemblies) with their correspondence with the final version of the transcriptome (final assembly, 4th). NT= not tested in qRT-PCR.'?' = the correspondence was not identified.

1st, 2nd and 3rd assemblies	4th assembly	Differentially expressed according to final assembly	Differentially expressed according to qRT-PCR
Locus1771v1rpkm127.15	comp29861_c0_seq1	YES	YES
Locus1124v1rpkm204.78	comp28026_c0_seq2	NO	NO
Locus11029v1rpkm8.14	comp23156_c0_seq2	YES	NO
Locus11310v1rpkm7.58	comp32365_c0_seq5	NO	NO
Locus13938v1rpkm28.95	comp32093_c0_seq8	NO	NT
Locus20443v1rpkm15.38, comp55263_c0_seq12, comp55263_c0_seq2, comp55424_c0_seq2	?	?	NO
Locus21788v1rpkm13.42	comp21967_c0_seq1, comp27269_c0_seq1	NO	NO
Locus61403v1rpkm0.00	comp28474_c0_seq2	NO	NT
Locus26783v1rpkm7.51	comp6261_c0_seq1	YES	NO
Locus26002v1rpkm8.34	comp24306_c0_seq2	YES	NT
Locus5941v5rpkm 5.03_PRE	comp41914_c0_seq1, comp25607_c0_seq1	NO	NT
Locus32024v1rpkm3.06	comp26319_c0_seq1	NO	NT
Locus27553v1rpkm6.74	comp32331_c0_seq2	NO	NO
Locus26972v1rpkm7.30	comp24462_c0_seq2	YES	NO
Locus31715v1rpkm3.28	comp31467_c0_seq1	NO	NT
Locus31370v1rpkm3.52	comp31640_c0_seq2	NO	NT
Locus31252v1rpkm3.60	comp25366_c0_seq1	NO	NT
Locus25079v1rpkm9.31	comp28591_c0_seq1	NO	NO

Locus52839v1rpkm0.00	comp13283_c0_seq1	YES	YES
comp55282_c0_seq1, comp55282_c0_seq3, comp54598_c0_seq1	comp9257_c0_seq1	NO	NT
comp55637_c0_seq16	comp27030_c0_seq1, comp32504_c0_seq1	NO	NO
comp53977_c1_seq12	comp29885_c0_seq8	NO	NO
comp55333_c3_seq3	comp31993_c0_seq1, comp31088_c0_seq1	NO	NO
comp42832_c0_seq2	comp25098_c0_seq2	NO	NT
comp43946_c0_seq5	comp23313_c0_seq1	NO	NT
comp53751_c0_seq1	comp30127_c0_seq2, comp26386_c0_seq1	NO	NT
comp55844_c1_seq5	comp32257_c0_seq1	NO	NT
Locus11230v1rpkm38.67, comp46228_c0_seq3	comp26595_c0_seq1	YES	YES
Locus27172v1rpkm7.10, comp47507_c2_seq2	comp28108_c0_seq1	YES	YES
	comp22480_c0_seq3	YES	NO
	comp27491_c0_seq3.1	YES	NO
	comp27022_c0_seq1.1	YES	NO
	comp25269_c0_seq1	YES	NO
	comp29120_c0_seq2.1	YES	NO
	comp20279_c0_seq4.1	YES	YES
	comp31481_c0_seq1.1	YES	NO
	comp25070_c0_seq2.1,	YES	NT

	comp17886_c0_seq1.1		
	comp6440_c0_seq1.1	YES	NT

After designing 63 primer pairs, a PCR on genomic DNA and a RT-PCR were performed to verify that the amplicon size was the one expected (specificity of amplification) and to confirm expression (i.e. that the transcript was not too poorly expressed) and evaluate whether a difference in intensity between bands from MT+ and MT- could be detected. The transcripts that gave robust and promising results in RT-PCR were then carried forward to the qRT-PCR analysis to obtain quantitative data and validate the results obtained by the *in silico* analysis.

To verify whether differences in gene expression were dependent on the mating type rather than on strain-specific characteristics, I performed qRT-PCR analysis on four couples of unrelated strains belonging to the more homogenous second set of samples (Table 2.3). Many of the transcripts belonging to the old assemblies resulted not differentially expressed according to mating type when tested by qRT-PCR, agreeing also with the result of the *in silico* differential expression analysis performed on the final assembly.

2.3.5 PCR and qRT-PCR validations

From now on, I will refer only to the last and definitive assembly (the 4th), resulting in 91 differentially expressed genes. Of these, 51 resulted up-regulated in the MT- samples while 40 in the MT+ ones. The list of differentially expressed transcripts is reported in APPENDIX A. All the genes previously validated as DEG according to mating type derived from the previous assemblies were present in the final assembly. However, only 17 of the 47 selected transcripts reported in Table 2.8 resulted predicted to be DEG in the final assembly. The results of the differential expression analysis for these 17 genes are reported

in Table 2.9, while the information concerning annotation and protein description are reported in Table 2.10.

Table 2.9: List of the transcripts selected from the differential expression analysis with the normalized counts provided for S1+ = Sy373 small, S2+ = B856 small, L2+ = B856 large, S1- = Sy379 small, S2- = B857 small, L2- = B857 large. LogFC= 2log fold change, Pvalue = p-value and FDR= False discovery rate.

Transcript ID	logFC	PValue	FDR	S1-	S2-	L2-	S1+	S2+	L2+
comp13283_c0_seq1	-10.44	7.98E-08	1.88E-04	0.70	0.70	0.17	1068.26	1621.92	0.75
comp29861_c0_seq1	-3.61	1.28E-04	4.47E-02	2.66	13.76	9.31	144.86	151.75	15.74
comp24462_c0_seq2	-13.28	1.90E-14	1.94E-10	0.00	0.00	0.00	22.89	7.79	7.13
comp22480_c0_seq3	-11.63	6.81E-22	2.09E-17	0.00	0.02	0.00	19.97	33.43	30.57
comp23156_c0_seq2	4.67	1.57E-04	5.34E-02	32.50	1.24	4.08	0.48	1.08	0.79
comp6261_c0_seq1	12.64	2.47E-13	1.30E-09	4.76	5.33	12.82	0.00	0.00	0.00
comp24306_c0_seq2	6.11	1.35E-05	8.47E-03	4.48	30.25	11.51	0.61	0.06	0.00
comp26595_c0_seq1	9.76	8.68E-07	9.52E-04	40.76	202.51	0.14	0.13	0.17	0.00
comp28108_c0_seq1	10.43	6.25E-07	7.99E-04	3.50	26.67	0.05	0.00	0.02	0.00
comp27491_c0_seq3	3.93	3.12E-04	8.33E-02	168.2	11.71	29.66	4.39	5.15	12.44
comp27022_c0_seq1	5.50	1.91E-05	1.05E-02	71.35	0.89	18.31	1.20	0.47	1.45
comp25269_c0_seq1	4.35	3.34E-07	5.40E-04	67.71	14.45	47.28	2.98	2.17	3.33
comp29120_c0_seq2	3.09	4.27E-06	3.28E-03	24.16	43.27	68.58	3.84	5.17	8.55
comp20279_c0_seq4	-12.17	3.91E-06	3.20E-03	0.00	0.00	0.00	9.89	7.14	0.00
comp31481_c0_seq1	-2.85	3.76E-04	9.46E-02	2.47	4.63	1.25	12.01	23.82	38.13
comp25070_c0_seq2, comp17886_c0_seq1	-7.14	1.64E-06	1.63E-03	0.00	0.06	0.00	2.27	3.87	3.91
comp6440_c0_seq1	7.75	1.20E-04	4.24E-02	0.84	29.96	0.00	0.06	0.06	0.02

Table 2.10: Transcripts annotation with the protein description of SwissProt (SP) and Conserved Domain; - = unknown.

Transcript ID	Description SP	Description CD
comp13283_c0_seq1	-	-
comp29861_c0_seq1	LRR receptor-like serine/threonine-protein kinase GSO1	PLN00113 leucine-rich repeat receptor-like protein kinase
comp24462_c0_seq2	-	-
comp22480_c0_seq3	-	pfam03151 TPT Triose-phosphate Transporter family
comp23156_c0_seq2	Heat shock factor protein	pfam00447 HSF_DNA-bind HSF-type DNA-binding
comp6261_c0_seq1	-	TIGR01444 fkbM_fam methyltransferase FkbM family
comp24306_c0_seq2	-	-
comp26595_c0_seq1	Probable leucine-rich repeat receptor-like protein kinase At1g35710	PLN00113 leucine-rich repeat receptor-like protein kinase
comp28108_c0_seq1	Heat shock factor protein 3	pfam00447 HSF_DNA-bind HSF-type DNA-binding
comp27491_c0_seq3	Putative oxidoreductase YteT	-
comp27022_c0_seq1	-	-
comp25269_c0_seq1	-	-
comp29120_c0_seq2	-	-
comp20279_c0_seq4	-	-
comp31481_c0_seq1	-	-
comp25070_c0_seq2, comp17886_c0_seq1	-	-
comp6440_c0_seq1	-	-

All the 17 transcripts were tested in RT-PCR; those that gave robust results (14 out of 17) were then carried forward to the qRT-PCR. This latter analysis proved that the differential expression was confirmed only for five transcripts, thus providing a group of genes related to mating types. All the others are false positive produced by the differential expression analysis probably due to the low number of replicas used to produce the RNA-seq. The MT-biased transcript ID, the assigned gene name and their logarithmic base2 fold change in qRT-PCR, compared to the FC in RNA-Seq are reported in Table 2.11.

Table 2.11: MT-related transcript ID, the assigned gene name and their logarithmic base2 fold change in qRT-PCR (mean and variance) compared to the FC in RNA-Seq.

Transcript ID	Gene name	Log2 FC qPCR	Log2 FC RNA-Seq
comp13283_c0_seq1	***MRP1***	9.8 ± 3	-10.44
comp29861_c0_seq1	***MRP2***	3.7 ± 2.6	-3.61
comp20279_c0_seq4	***MRP3***	7.5± 5	-12.17
comp28108_c0_seq1	***MRM1***	6.6 ± 0.8	10.43
comp26595_c0_seq1	***MRM2***	9.7 ± 4.6	9.76

The acronyms MRP and MRM stand, respectively, for **M**ating type **R**elated **P**lus and **M**ating type **R**elated **M**inus. *MRP1, MRP2* and *MRP3* resulted up to 12.7 folds more expressed in MT+ samples as compared to MT- ones, while *MRM1* and *MRM2* resulted up to 8.4 folds more expressed in MT- samples as compared to MT+ ones (Figs 2.2, 2.3, 2.4, 2.5, 2.6). The expression variation was calculated setting one of the four MT- or MT+ as control condition against the other MT+ or MT- samples, and was normalized over the expression variation of reference genes whose expression levels were not regulated in these specific experimental conditions. The relative expression ratio (R) of the targeted mating-type related genes was computed separately for each of the four samples settled as controls (data not shown) to verify that the result was not changing in relation to the sample chosen as reference condition.

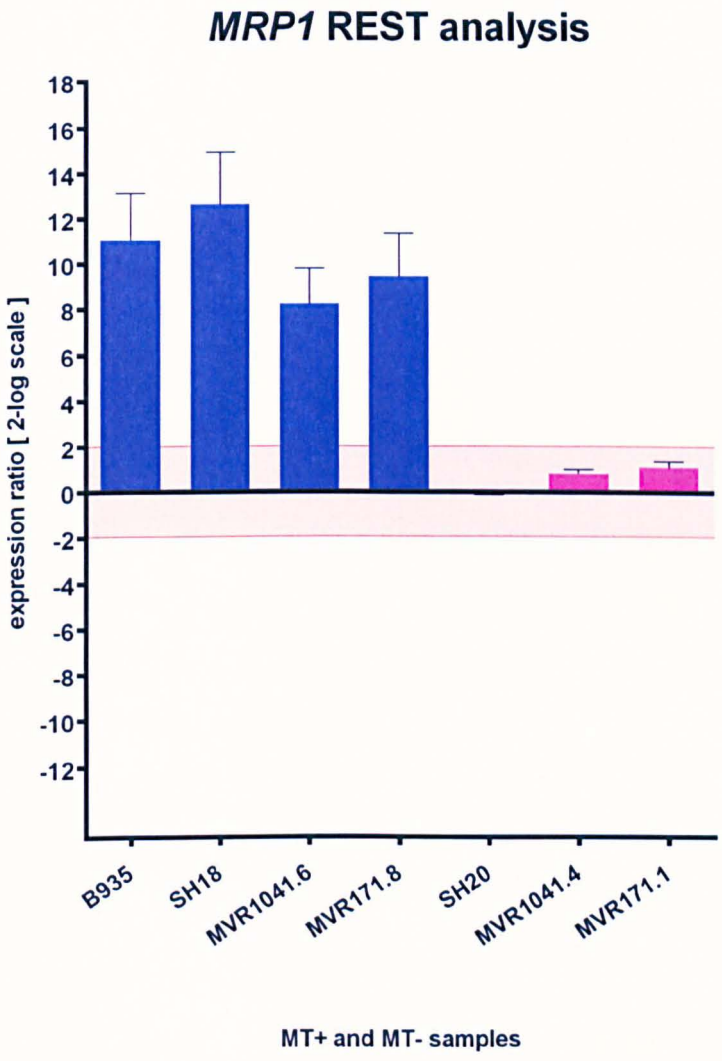


Figure 2.2: REST analysis of *MRP1*. Reference condition: B936 MT-, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

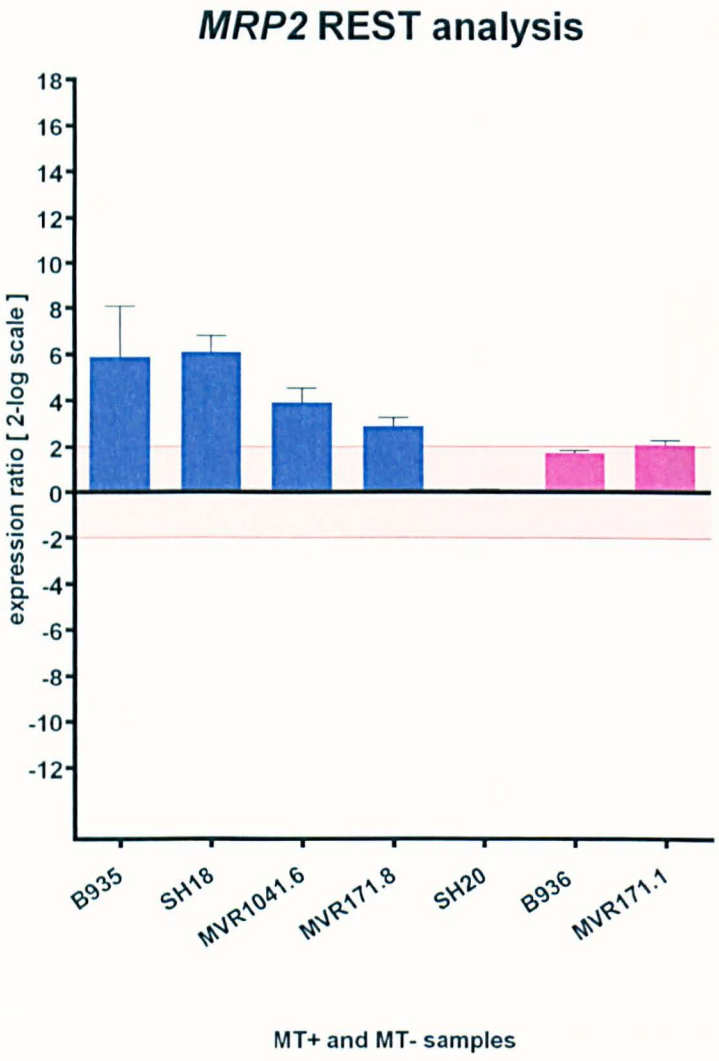


Figure 2.3: REST analysis of *MRP2*. Reference condition: MVR1041.4 MT-, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

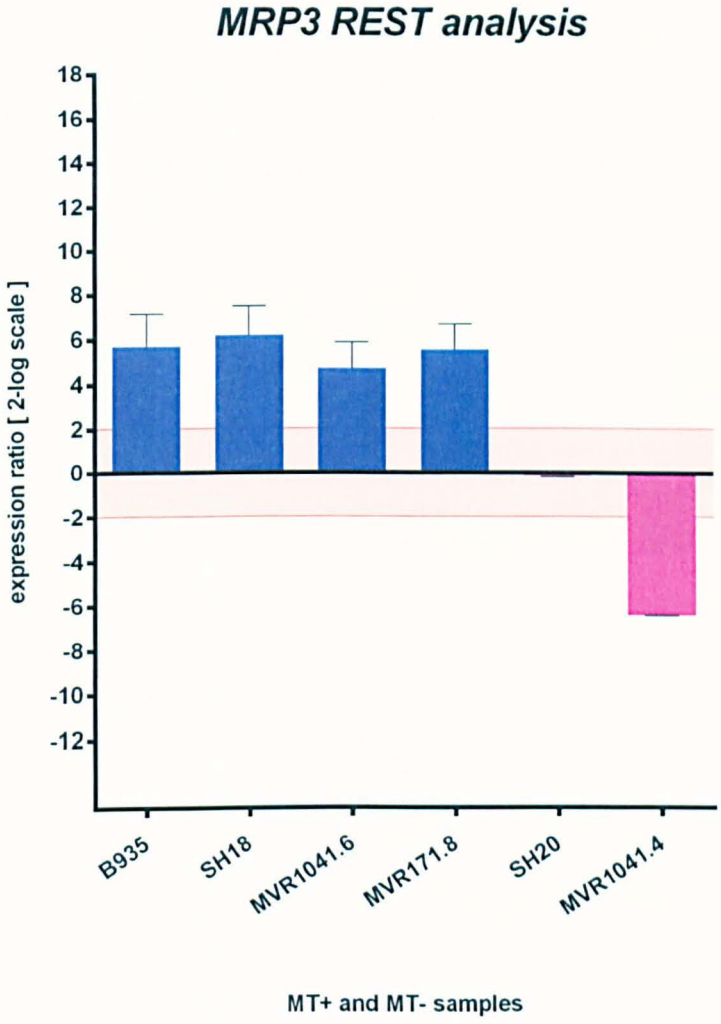


Figure 2.4: REST analysis of *MRP3*. Reference condition: MVR171.1 MT-, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

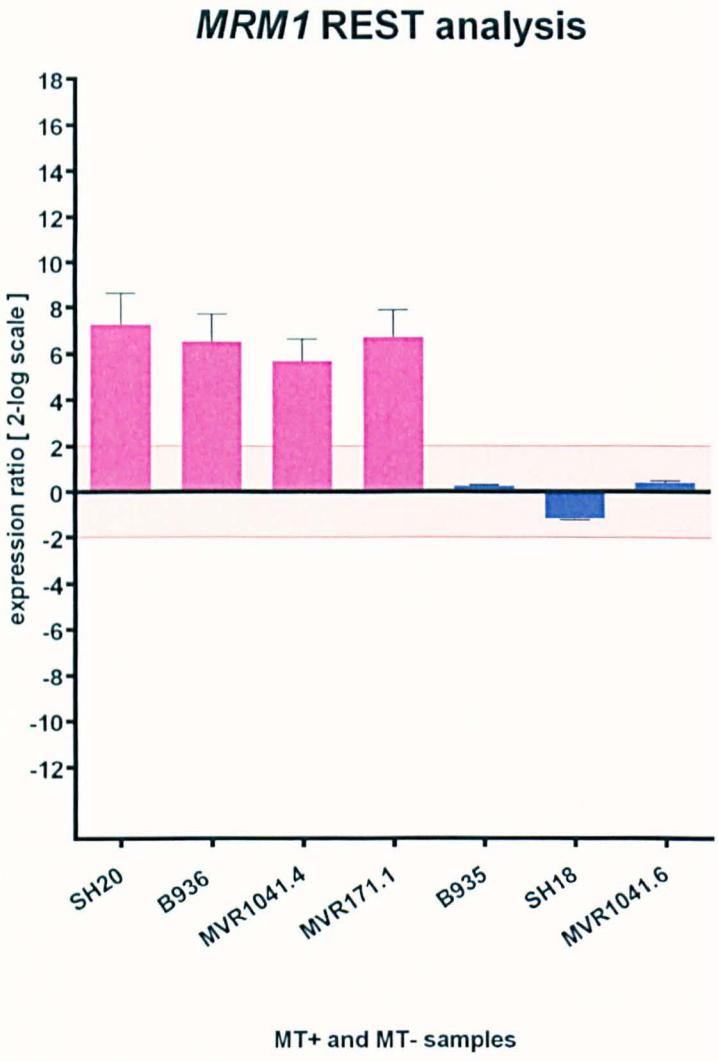


Figure 2.5: REST analysis of *MRM1*. Reference condition: MVR171.8 MT+, reference genes: *CDK*, *TUB A* and TUB *B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

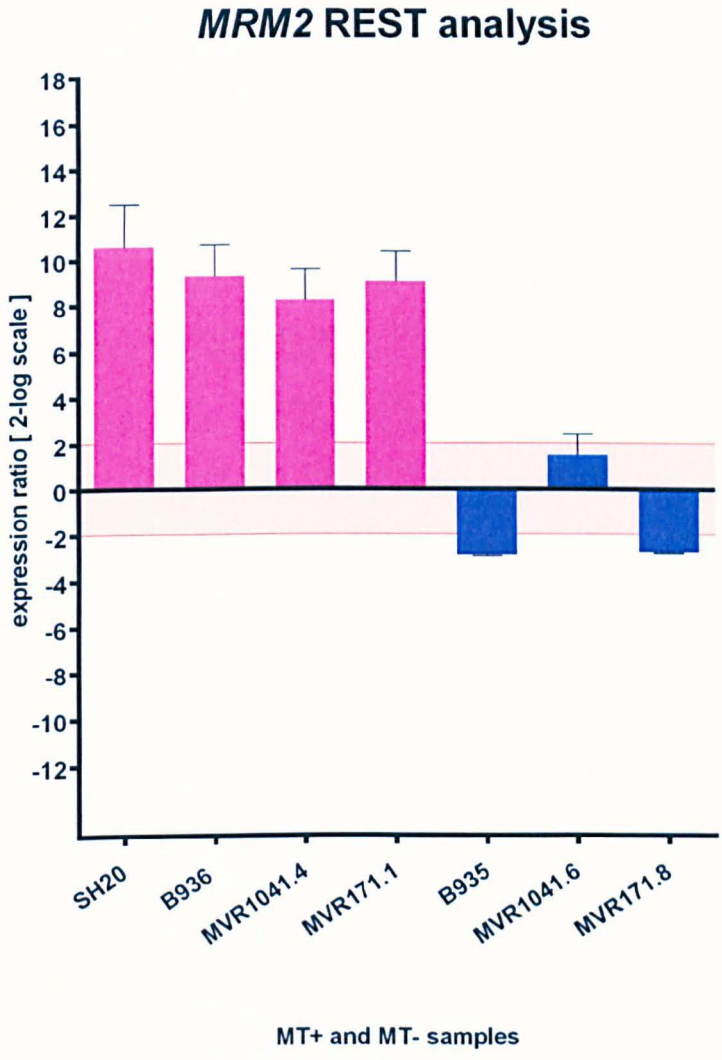


Figure 2.6: REST analysis of *MRM2*. Reference condition: SH18 MT+, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

Out of the five transcripts resulting differentially expressed according to mating type, the REST analyses of the remaining nine transcripts showed that in four of them the differential expression was not related to the mating type but to the strain-specific variability of the samples of *P. multistriata*. Five transcripts resulted not differentially expressed. REST analyses of the nine transcripts that were not differentially expressed according to mating type are reported in APPENDIX B.

2.3.6 Characterization of the five MT-biased genes

MRP1 (Mating type Related Plus1) is a MT+ biased gene of *Pseudo-nitzschia multistriata.* It is located on PsnmuV1.4_scaffold_157 -size_117502:94841..95852 of the *P. multistriata* genome (-strand) and contains one intron of 112 bp. The predicted gene model, PSNMU-V1.4_AUG-EV-PASAV3_0024820.1, was edited thanks to the available RNA-seq tracks that were showing the correct gene structure as comp13283_c0_seq1.

MRP1 transcript is 901 bp long, including the 5'-3'UTR, and the 597 bp ORF, on reading frame +3 (RF+3), encodes for a 199 amino acids protein of unknown function. The protein has a predicted molecular weight of 21978.57 Daltons. SMART identified a signal peptide of 22 AA (MMTFNFSTVVLALVAATSFVSA) with its cleavage site between positions 22 and 23 (VSA-DY) in its protein sequence at N-terminus, and two low complexity regions. Both SMART and WOLF PSORT predicted that the protein region following the signal peptide presents a non-cytoplasmic domain with a probable extracellular localization. Moreover, ASAFind version 1.1.6 detected that the protein was not plastid localized.

MRP2 (Mating type Related Plus2) is the second MT+ biased gene validated. It is located on PsnmuV1.4_scaffold_91-size_164462:94087..96315 (- strand) of the *P. multistriata* genome, is named PSNMU-V1.4_AUG-EV-PASAV3_0122240.1 and contains one intron of 162 bp confirmed by PCR results on genomic DNA (data not shown).

MRP2 transcript is 2122 bp long and the resulting ORF (RF+1) is 1842 bp long. The 613 AA protein was predicted to be a leucine-rich repeat (LRR) receptor-like serine/threonine-protein kinase GSO1, with a predicted molecular weight of 68341.16 Daltons. However, further analysis proved that the protein structure had no recognisable serine/threonine-protein kinase as predicted by Annocript. This is demonstrated by the alignment presented in Figure 2.7. The protein sequence of *MRP2* was aligned with two best matches in the

NCBI protein sequence database annotated as leucine-rich repeat receptor-like protein kinase.

SignalP did not detect any signal peptide, NCBI_CD detected a multi-domain of LRR_8 (Leucin Rich Repeat 8) while SMART identified a transmembrane region before the LRR one. WOLF PSORT predicted that the protein had a probable localization in the endoplasmic reticulum; however, it is important to consider that the result is not optimized for diatoms. Since no kinase domain and signal peptide were detected, *MRP2* has to be redefined as probable leucine-rich repeat-containing protein.

A)

Accession	Description
XP_003554177.1	PREDICTED: leucine-rich repeat receptor-like tyrosine-protein kinase PXC3 [Glycine max]
NP_173166.2	leucine-rich receptor-like protein kinase [Arabidopsis thaliana]
lcl|Query_10001	MRP2

B)

```
ref|XP_003554177.1|  ------------------------------------------------------------
ref|NP_173166.2|     ------------------------------------------------------------
lcl|Query_10001      MSLNNQANPCTVVAEVIDPNAVVTAPRYPRARPYSQRHLDGMEAQVVNQEQSQHVTFEAD

ref|XP_003554177.1|  ------------------------------------------------------------
ref|NP_173166.2|     ----------------------------MRGRICFLAIVILCSFSFILVRSLNEEGRVLL
lcl|Query_10001      NKAAFFADSSEGKNRFGKVSNEKKGYNLLSSKWYWYAMIAMTLIFLVLVGYGFGSGLFLL

ref|XP_003554177.1|  ------------------------------------------------------------
ref|NP_173166.2|     EFKAFLNDSNGYLASWNQLDSNPCNWTGIACTHLRTVTSVDLNGMNLSGTLSPLICKLHG
lcl|Query_10001      K------RGNATLNNDPAKDGSTPNGSDPSSISDSSNTVQDRQDYKYS------IVTLLG

ref|XP_003554177.1|  ------------------------------------------------------------
ref|NP_173166.2|     LRKLNVSTNFISGPIPQDLSLCRSLEVLDLCTNRFHGVIPIQLTMIITLKKLYLCENYLF
lcl|Query_10001      LPMVMERTS------PQ----ARAVEWL----------------------AYQDEPLF-

ref|XP_003554177.1|  ----------------------------MEFVCLLYILLAWCLSSSELVGAELQDQDI
ref|NP_173166.2|     GSIPRQIGNLSSLQELVIYSNNLTGVIPPSMAKLRQLRIIRAGRNGFSGVIPSEISGCES
lcl|Query_10001      -----DVTKETSAEEQDRHH----------EILEQRYALVVWYF--------DQGGPTV

ref|XP_003554177.1|  LNAINQ-----ELRVPGWGDA--NNSNYCTWQG-------VSCGNHSMVEGLDLSHRNLR
ref|NP_173166.2|     LKVLGLAENLLEGSLPKQLEKLQNLTDLILWQNRLSGEIPPSVGNISRLEVLALHENYFT
lcl|Query_10001      WKTINRDESA------GWVEF-GAGVHECDWKGV---DCDYENGNTETGTVIGLRLSPAL

ref|XP_003554177.1|  G------NVTL-MSELKALKRLDLSNNNFDGSIPPAFGNLSDLEVLDLSSNKFQGSIPP-
ref|NP_173166.2|     G------SIPREIGKLTKMKRLYLYTNQLTGEIPREIGNLIDAAEIDFSENQLTGFIPK-
lcl|Query_10001      GLVLTGTHLSTELGMLTALRRVDFSDQRLQGTIPDEWASMTNLKSVILSKNQLQTTVPEW

ref|XP_003554177.1|  ------QLGGLTNLKSLNLSNNVLVGEIPIELQGLEKLQDFQISSNHLSGLVPSWVGNLT
ref|NP_173166.2|     ------EFGHILNLKLLHLFENILLGPIPRELGELTLLEKLDLSINRLNGTIPQELQFLP
lcl|Query_10001      IGRSSEEGGGWQNLEQLALDDNFLYGSLPSSMRNLRKLTHLELQVN------PQLEGRFD

ref|XP_003554177.1|  NLRLFTAYENR--LDGRIPDDLGLISDLQILNLHSNQLEGPIPASIFVPGKLEVLVLTQN
ref|NP_173166.2|     YLVDLQLFDNQ--LEGKIPPLIGFYSNFSVLDMSANSLSGPIPAHFCRFQTLILLSLGSN
lcl|Query_10001      ELLFLEKEDEATGLPGQ---------NLEILDLSNTRLKGKIP-KITLPSIRSFRLWNLP
```

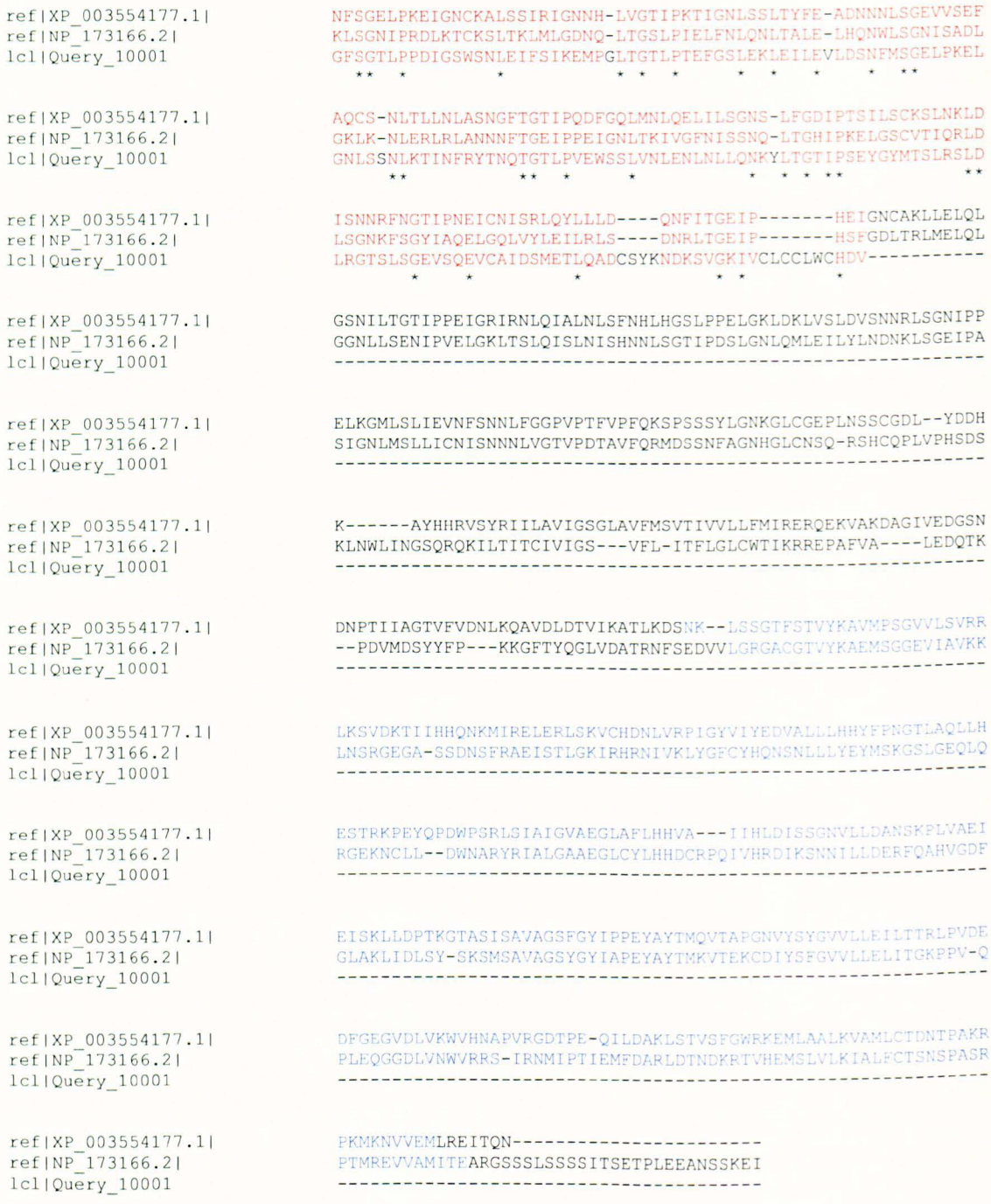


Figure 2.7: CLUSTAL W multiple sequence alignment. A. The accession number of the two best matches in the NCBI protein sequence database annotated as leucine-rich repeat receptor-like tyrosine-protein kinase. B. The alignment with highlighted the conserved a.a. (*) and the protein domains (green: transmembrane domain of MRP2, red: conserved leucin rich repeats, red bold: LRR_8 domain, blue: Serine/Threonine protein kinases absent from MRP2).

MRP3 (Mating-type Related Plus 3) resulted to be a MT+ biased gene. Its correspondent gene model is PSNMU-V1.4_AUG-EV-PASAV3_0020770.1 present on

PsnmuV1.4_scaffold_147-size_134965:92696..93904 (+ strand). It is 1209 bp long and its ORF, of 828 bp, encodes for a 276 amino acids protein of unknown function with molecular weight of 31340.27 Daltons. WOLF PSORT predicted that the protein had a probable nuclear/cytosolic localization; however, it is important to consider that the result is not optimized for diatoms. No other information was achieved from the BLAST analyses described in Chapter 2.2.11.

MRM1 (Mating-type Related Minus1) is a MT- biased gene. It was found twice in the *P. multistriata* genome: on PsnmuV1.4_scaffold_432-size_35274:4076..5639 (+ strand), named PSNMU-V1.4_AUG-EV-PASAV3_0085380.1, and on PsnmuV1.4_scaffold_204-size_91430:87746..89292 (+ strand) where it was named PSNMU-V1.4_AUG-EV-PASAV3_0041130.1. The two scaffolds partially overlap (almost 8 kb at the end of scaffold_204 and at the beginning of scaffold_432), however they are not identical due an insertion of 102 bp on scaffold_432 at position 6183..6285. The possible explanation for such scenario could be a duplication or a high polymorphic nature of the genomic region. PCR analyses followed by Sanger sequencing were performed to obtain more insights into this region. To start with, I performed PCR and sequencing on the 5' flanking region of the *MRM1* gene at the level of the putative promoter, and identified a 5 bp in/del. The in/del displays three possible alternatives: 1) 5 bp deletion in homozygosis; 2) no deletion in homozygosis; 3) deletion in heterozygosis (Fig. 2.8). Three MT+ and four MT- were analysed and no specific pattern for the absence/presence of the deletion could be found: MT- strains could display any of the three possible genotypes, indicating that this polymorphism does not correlate with the MT. Further PCR and sequencing analyses are on-going to explore the region.

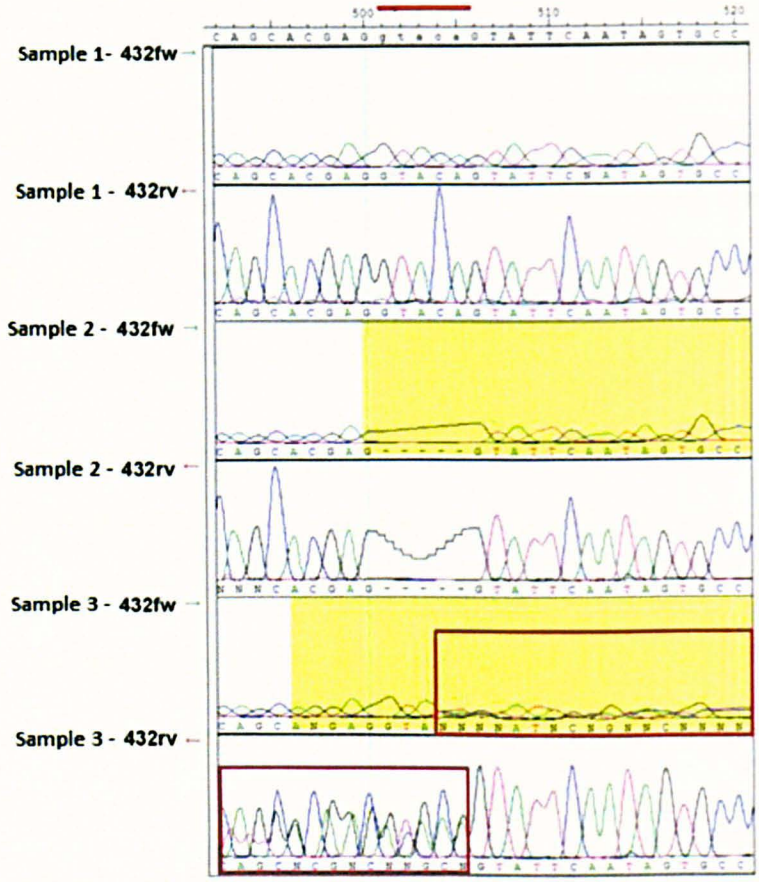


Figure 2.8: A 5bp in/del present in the promoter region of *MRM1*. Electropherograms showing sequences of the putative promoter region (-486/-457 bp upstream of the putative start site) in three samples. The first, third and fifth sequences were obtained with a forward primer (green arrow) while the second, fourth and sixth sequences were obtained with a reverse primer (red arrow). The first four sequences shows that the in/del (GTACA) (marked with a red bar on the consensus sequence) could be present (first and second) or absent (third and fourth). The fifth and sixth sequences display double peaks (boxed in red) indicating that the in/del is in heterozygosis.

MRM1 contained one intron of 128 bp. Its transcript, comp20108_c0_seq1, was 1546 bp long. The ORF (RF+1) was 1221 bp long encoding for 406 AA annotated as Heat Shock Factor (HSF)-type DNA-binding domain protein 3. The protein had a predicted molecular weight of 44311.29 Daltons. WOLF PSORT predicted that the protein had a probable nuclear localization; however, it is important to consider that the result is not optimized for diatoms.

MRM2 (Mating-type Related Minus2) is a MT- biased gene. It was located on PsnmuV1.4_scaffold_11-size_341144:203626..205675 (+ strand) of the genome, named

PSNMU-V1.4_AUG-EV-PASAV3_0006960.1. It was 2022 bp long and contained two introns, one of 96 bp and one of 120 bp, confirmed by PCR results on genomic DNA (data not shown). *MRM2* (ORF in RF+3) encodes for a 539 AA protein annotated as Leucine-rich repeat receptor-like protein kinase. The leucin rich repeat domains were found in between position 208..522 and a transmembrane region was detected before the LRR domain. However, further analysis proved that the protein structure had no recognisable protein kinase, as predicted by Annocript. This is demonstrated by the alignment presented in Figure 2.9. The protein sequence of *MRM2* was aligned with two best matches in the NCBI protein sequence database annotated as leucine-rich repeat receptor-like protein kinase.

MRM2 protein had a predicted molecular weight of 59773.06 Daltons. WOLF PSORT predicted that the protein had a probable nuclear/cytosolic localization, however is important to consider that the result is not optimized for diatoms. Revision of the protein annotation brought to hypothesize at a probable Leucin rich repeat (LRR)-containing protein.

A)

Accession	Description
lcl|1	lcl|1 MRM2
gi|955385922	PREDICTED: probable LRR receptor-like serine/threonine-protein kinase At4g08850 isoform X1 [*Glycine max*]
gi|15219370	leucine-rich repeat receptor-like protein kinase PEPR1 [*Arabidopsis thaliana*]

B)

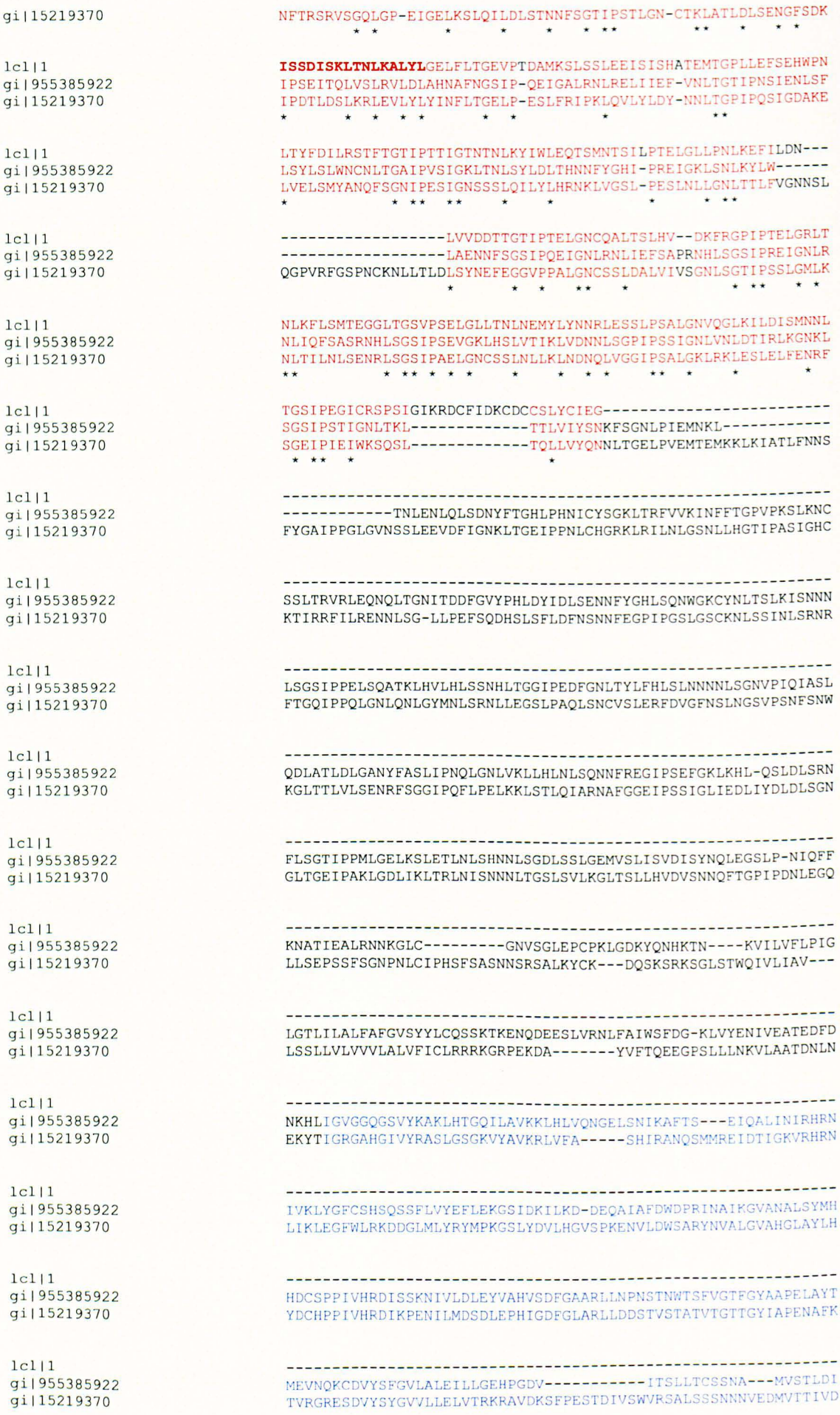
```
lcl|1          MLATSDPYPSPHAKPEANVEYGNVQRFPVESDDGYRAQDQRKNRNSYRPLFITLYVMLLL
gi|955385922   ------------------------------------------------------------
gi|15219370    ------------------------------------------------------------

lcl|1          ATAGLAFAVTRYVQTVKNKHKSPPSQDTTSDGSTDVDSSSPATLAEVDPDRDLIAYRSDI
gi|955385922   ----------------------------------------------------MHKLHTNL
gi|15219370    ----------------------------------------------------MK----NL

lcl|1          EYILFTEMDEGLSTDFVEGPQKRAIDWLVSDDL------VLNSTEVRAMAEYIKNGDEDS
gi|955385922   GNHIFDCIRQSLIVEYPIPPMMFIFPTLQSMKLPSFWLLLIVMLFCAFTMATSRHATIPS
gi|15219370    GG-LFK-----------------------------------ILLLFFCLFLSTHIISVSCLN
                   *

lcl|1          VSTVPLVQRYALMVLFFATNGELWSD---SSWRELVNVPE---CRFMGIECDLEGHINTL
gi|955385922   SASLTLQQTEANALLKWKTSLDNQSQALLSSWG--GN----TPCNWLGIACDHTKSVSSI
gi|15219370    SDGLTLL-----SLLK---HLDRVPPQVTSTWK--INASEATPCNWFGITCDDSKNVASL
                   *       *           * *   *     *   ** **

lcl|1          DVGYRKLRGRLPG-EVGMLSMLTSLNVESNNLEGTIPSFLYNKLTKLERVDMRNNGFLST
gi|955385922   NLTHVGLSGMLQTLNFSSLPNILTLDMSNNSLKGSIPPQIRV-LSKLTHLDLSDNHFSGQ
```


gi|15219370 NFTRSRVSGQLGP-EIGELKSLQILDLSTNNFSGTIPSTLGN-CTKLATLDLSENGFSDK
lcl|1 ISSDISKLTNLKALYLGELFLTGEVPTDAMKSLSSLEEISISHATEMTGPLLEFSEHWPN
gi|955385922 IPSEITQLVSLRVLDLAHNAFNGSIP-QEIGALRNLRELIIEF-VNLTGTIPNSIENLSF
gi|15219370 IPDTLDSLKRLEVLYLYINFLTGELP-ESLFRIPKLQVLYLDY-NNLTGPIPQSIGDAKE
lcl|1 LTYFDILRSTFTGTIPTTIGTNTNLKYIWLEQTSMNTSILPTELGLLPNLKEFILDN---
gi|955385922 LSYLSLWNCNLTGAIPVSIGKLTNLSYLDLTHNNFYGHI-PREIGKLSNLKYLW------
gi|15219370 LVELSMYANQFSGNIPESIGNSSSLQILYLHRNKLVGSL-PESLNLLGNLTTLFVGNNSL
lcl|1 ------------------LVVDDTTGTIPTELGNCQALTSLHV--DKFRGPIPTELGRLT
gi|955385922 ------------------LAENNFSGSIPQEIGNLRNLIEFSAPRNHLSGSIPREIGNLR
gi|15219370 QGPVRFGSPNCKNLLTLDLSYNEFEGGVPPALGNCSSLDALVIVSGNLSGTIPSSLGMLK
lcl|1 NLKFLSMTEGGLTGSVPSELGLLTNLNEMYLYNNRLESSLPSALGNVQGLKILDISMNNL
gi|955385922 NLIQFSASRNHLSGSIPSEVGKLHSLVTIKLVDNNLSGPIPSSIGNLVNLDTIRLKGNKL
gi|15219370 NLTILNLSENRLSGSIPAELGNCSSLNLLKLNDNQLVGGIPSALGKLRKLESLELFENRF
lcl|1 TGSIPEGICRSPSIGIKRDCFIDKCDCCSLYCIEG------------------------
gi|955385922 SGSIPSTIGNLTKL-------------TTLVIYSNKFSGNLPIEMNKL------------
gi|15219370 SGEIPIEIWKSQSL-------------TQLLVYQNNLTGELPVEMTEMKKLKIATLFNNS
lcl|1 --
gi|955385922 ------------TNLENLQLSDNYFTGHLPHNICYSGKLTRFVVKINFFTGPVPKSLKNC
gi|15219370 FYGAIPPGLGVNSSLEEVDFIGNKLTGEIPPNLCHGRKLRILNLGSNLLHGTIPASIGHC
lcl|1 --
gi|955385922 SSLTRVRLEQNQLTGNITDDFGVYPHLDYIDLSENNFYGHLSQNWGKCYNLTSLKISNNN
gi|15219370 KTIRRFILRENNLSG-LLPEFSQDHSLSFLDFNSNNFEGPIPGSLGSCKNLSSINLSRNR
lcl|1 --
gi|955385922 LSGSIPPELSQATKLHVLHLSSNHLTGGIPEDFGNLTYLFHLSLNNNNLSGNVPIQIASL
gi|15219370 FTGQIPPQLGNLQNLGYMNLSRNLLEGSLPAQLSNCVSLERFDVGFNSLNGSVPSNFSNW
lcl|1 --
gi|955385922 QDLATLDLGANYFASLIPNQLGNLVKLLHLNLSQNNFREGIPSEFGKLKHL-QSLDLSRN
gi|15219370 KGLTTLVLSENRFSGGIPQFLPELKKLSTLQIARNAFGGEIPSSIGLIEDLIYDLDLSGN
lcl|1 --
gi|955385922 FLSGTIPPMLGELKSLETLNLSHNNLSGDLSSLGEMVSLISVDISYNQLEGSLP-NIQFF
gi|15219370 GLTGEIPAKLGDLIKLTRLNISNNNLTGSLSVLKGLTSLLHVDVSNNQFTGPIPDNLEGQ
lcl|1 --
gi|955385922 KNATIEALRNNKGLC---------GNVSGLEPCPKLGDKYQNHKTN----KVILVFLPIG
gi|15219370 LLSEPSSFSGNPNLCIPHSFSASNNSRSALKYCK---DQSKSRKSGLSTWQIVLIAV---
lcl|1 --
gi|955385922 LGTLILALFAFGVSYYLCQSSKTKENQDEESLVRNLFAIWSFDG-KLVYENIVEATEDFD
gi|15219370 LSSLLVLVVVLALVFICLRRRKGRPEKDA-------YVFTQEEGPSLLLNKVLAATDNLN
lcl|1 --
gi|955385922 NKHLIGVGGQGSVYKAKLHTGQILAVKKLHLVQNGELSNIKAFTS---EIQALINIRHRN
gi|15219370 EKYTIGRGAHGIVYRASLGSGKVYAVKRLVFA-----SHIRANQSMMREIDTIGKVRHRN
lcl|1 --
gi|955385922 IVKLYGFCSHSQSSFLVYEFLEKGSIDKILKD-DEQAIAFDWDPRINAIKGVANALSYMH
gi|15219370 LIKLEGFWLRKDDGLMLYRYMPKGSLYDVLHGVSPKENVLDWSARYNVALGVAHGLAYLH
lcl|1 --
gi|955385922 HDCSPPIVHRDISSKNIVLDLEYVAHVSDFGAARLLNPNSTNWTSFVGTFGYAAPELAYT
gi|15219370 YDCHPPIVHRDIKPENILMDSDLEPHIGDFGLARLLDDSTVSTATVTGTTGYIAPENAFK
lcl|1 --
gi|955385922 MEVNQKCDVYSFGVLALEILLGEHPGDV-----------ITSLLTCSSNA---MVSTLDI
gi|15219370 TVRGRESDVYSYGVVLLELVTRKRAVDKSFPESTDIVSWVRSALSSSNNNVEDMVTTIVD

```
lcl|1                     ------------------------------------------------------------
gi|955385922              PSLMGKLDQRLPYPINQMAKEIALIAKTAIACLIESPHSRPTMEQVAKELGMSK------
gi|15219370               PILVDELLD------SSLREQVMQVTELALSCTQQDPAMRPTMRDAVKLLEDVKHLARSC

lcl|1                     ------
gi|955385922              -SSSVH
gi|15219370               SSDSVR
```

Figure 2.9: CLUSTAL W multiple sequence alignment. A. The accession number of the two best matches in the NCBI protein sequence database annotated as leucine-rich repeat receptor-like protein kinase. B. The alignment with highlighted the conserved a.a. (*) and protein domains (green: transmembrane domain of *MRM2*, red: conserved leucin rich repeats, red bold: LRR_8 domain, blue: Serine/Threonine kinases, Interleukin-1 Receptor Associated Kinases in *Glycine* and Protein Kinases, catalytic domain in *Arabidopsis*, absent from MRM2).

2.3.7 Conservation of the five MT-biased genes in other species

TBLASTN analyses were conducted to search for homologs of the five MT-biased genes in five diatom genomes and in the transcriptomes of the MMETSP project. Retrieved protein sequences were classified as homologs based on the significance (e-value) and on the percentage of sequence similarity, and homology was confirmed with a reciprocal TBLASTN on the *P. multistriata* genome and transcriptome.

The shared conservation of the five *P. multistriata* MT-biased genes with other diatom species is summarized in Figure 2.10. Homology was recorded with sequences of other *Pseudo-nitzschia* species, of the phylogenetically closely related *Fragilariopsis cylindrus* and *F. kerguelensis*, and of three strains of raphid diatoms (two *Nitzschia* and *Cylidrotheca closterium*). Homologues of gene *MRM2* were recorded also in two araphid pennates (*Staurosira* complex and *Cyclophora tenuis).* A homologue of gene *MRM1* was recorded also in the centric diatom *Skeletonema marinoi* and a homologue of *MRP2* in *Thalassiosira weissflogii.*

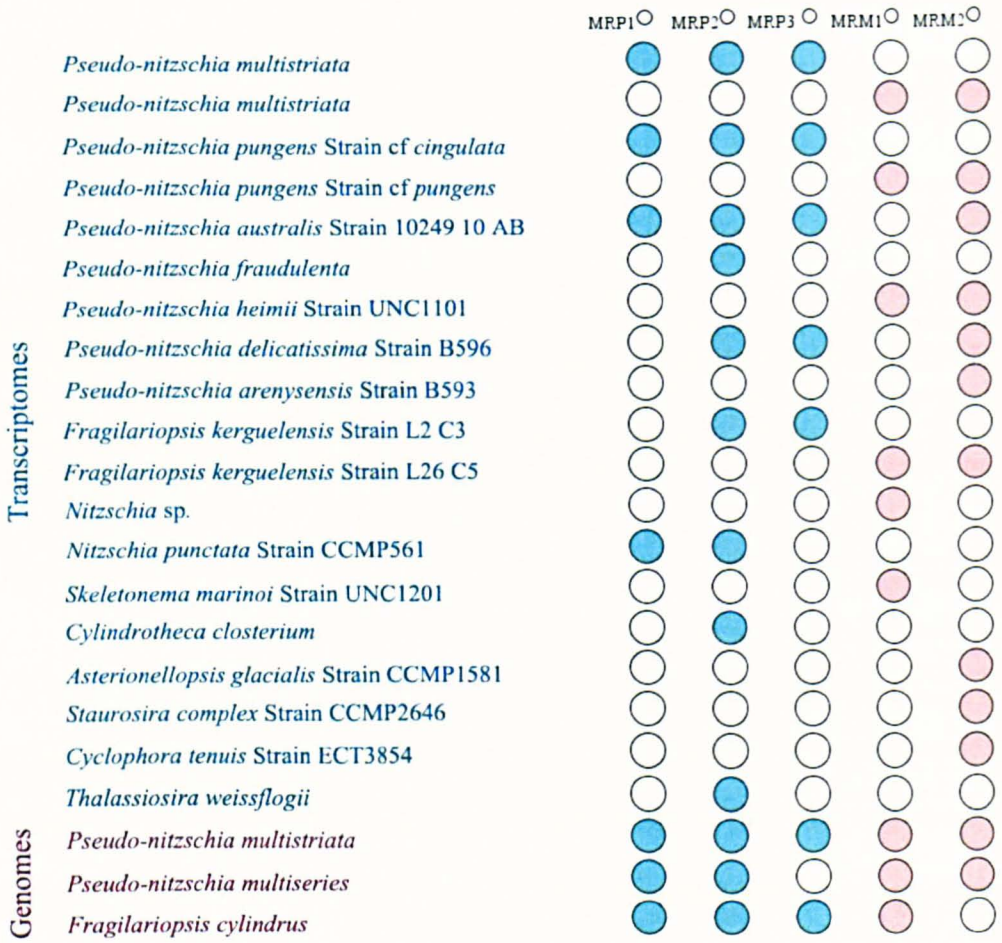


Figure 2.10: Coulson Plot (Field *et al.* 2013), graphical representation of the conservation of the five *P multistriata* MT-biased genes. The species are listed for both transcriptomes and genomes according to taxonomic relation. The species for which only the transcriptome was available are reported in blue, those for which the genome was available are reported in red. The filled circle highlights the presence of protein homology while the empty circles mean that no homologous proteins were present.

Interestingly, no homologous sequences could be found in the *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* genomes, nor in the *Seminavis robusta* genome (unpublished genome; courtesy of W. Vyverman), nor in the genome of *Ectocarpus siliculosus*, a multicellular Stramenopile.

Phylogenetic analyses were performed on the proteins classified as homologs to highlight how the different sequences relate in terms of sequence similarity. The ML (Maximum likelihood) method was used to create phylogenetic trees for the five genes and their homologs (Figs 2.11, , 2.12, 2.13, 2.14, 2.15).

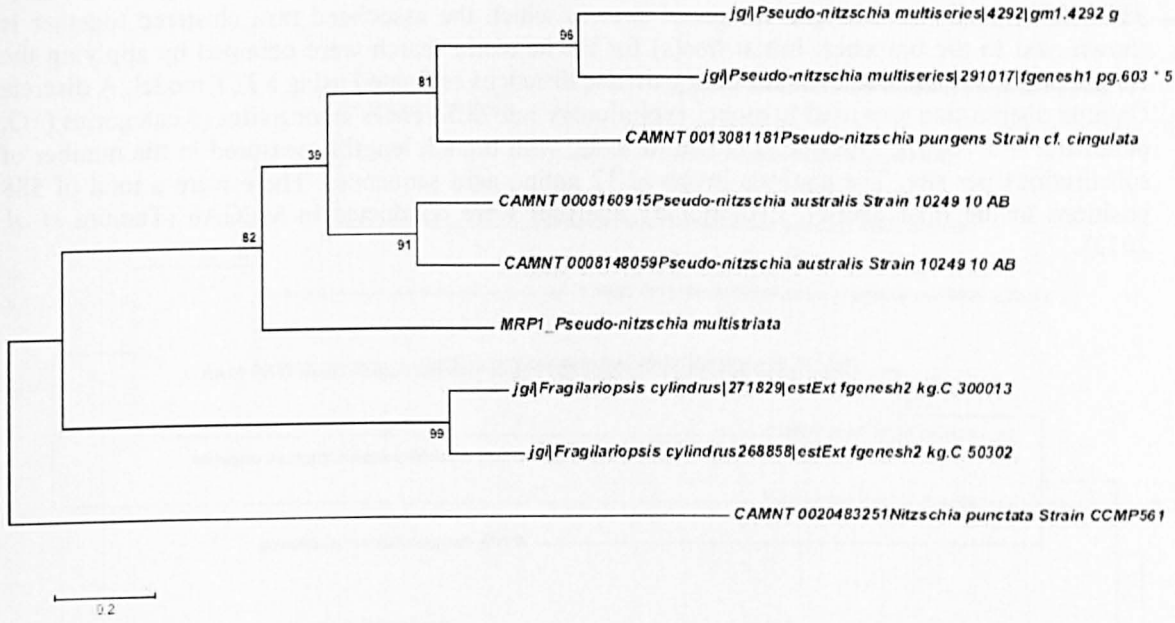


Figure 2.11: Molecular Phylogenetic analysis by Maximum Likelihood method of gene *MRP1*. The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan And Goldman model (Whelan and Goldman 2001). The tree with the highest log likelihood (-3399.4931) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.0439)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 9 amino acid sequences. There were a total of 230 positions in the final dataset. Phylogenetic analyses were conducted in MEGA6 (Tamura *et al.* 2013).

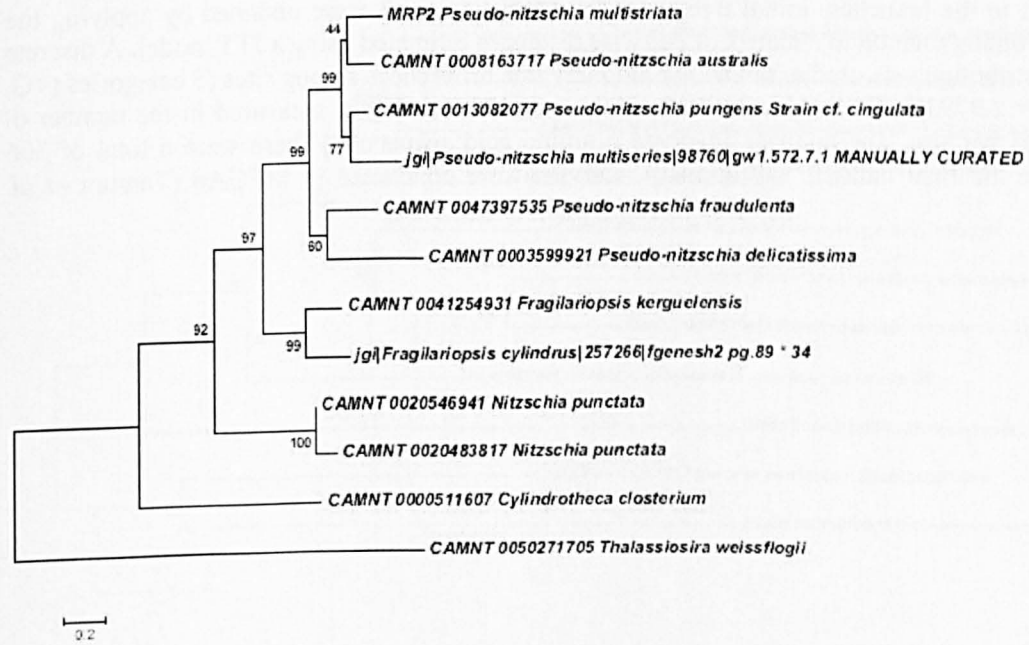


Figure 2.12: Molecular Phylogenetic analysis by Maximum Likelihood method of gene *MRP2*.The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan

And Goldman model (Whelan and Goldman 2001). The tree with the highest log likelihood (-9117.3831) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.7620)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 12 amino acid sequences. There were a total of 588 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.* 2013).

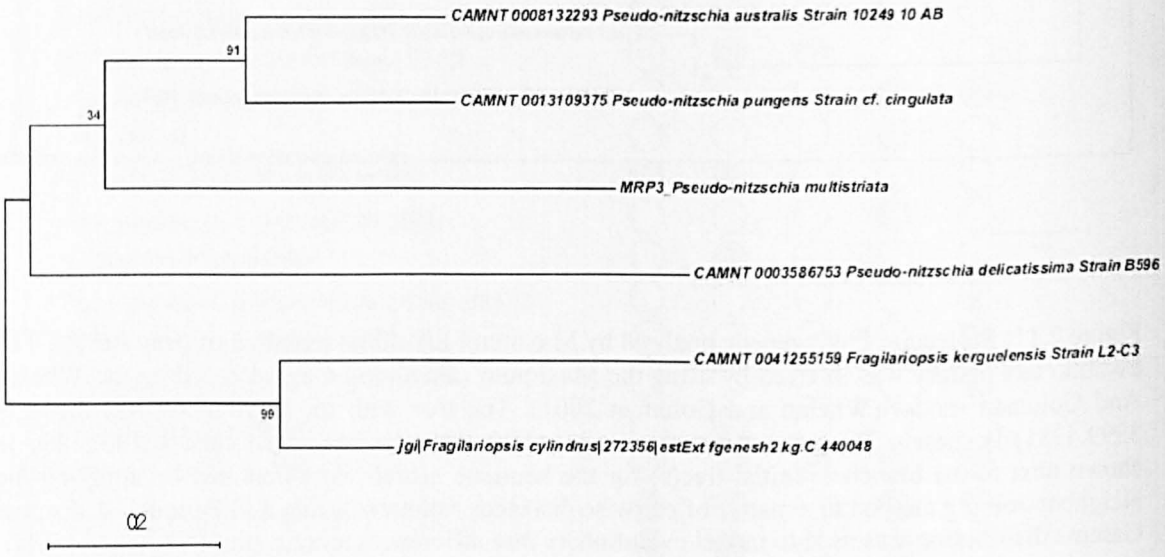


Figure 2.13: Molecular Phylogenetic analysis by Maximum Likelihood method of gene *MRP3*.The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model (Whelan and Goldman 2001). The tree with the highest log likelihood (-2933.0805) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.9793)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 6 amino acid sequences. There were a total of 306 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.* 2013).

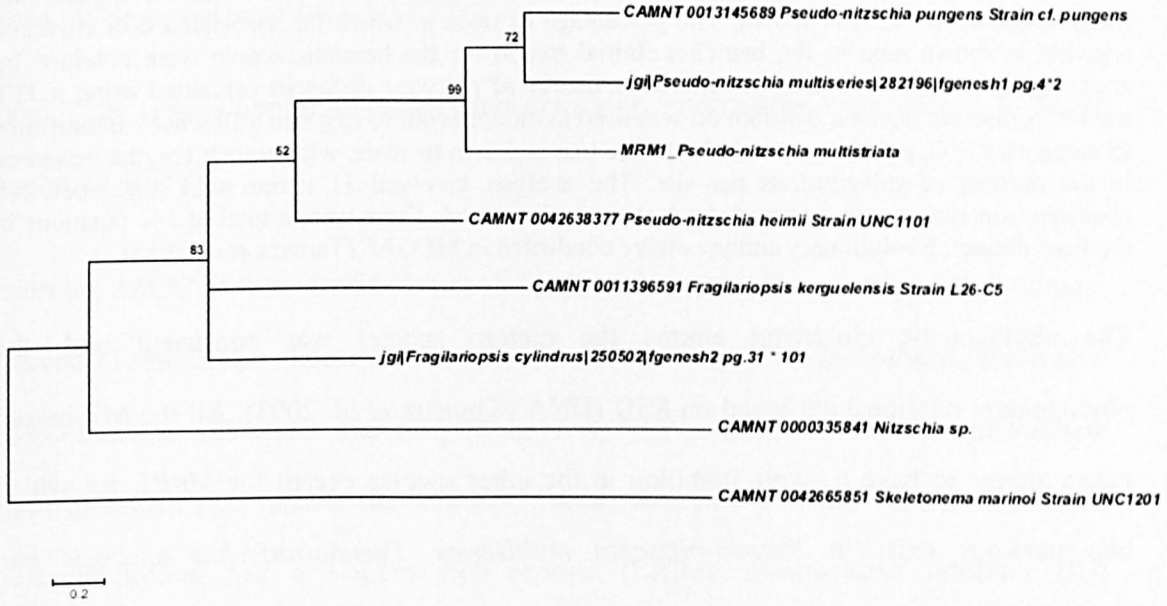


Figure 2.14: Molecular Phylogenetic analysis by Maximum Likelihood method of gene *MRM1*. The evolutionary history was inferred by using the Maximum Likelihood method based on the Jones et al. w/freq. model (Whelan and Goldman 2001). The tree with the highest log likelihood (-2912.0726) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.4828)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 0.0000% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 8 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 171 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.* 2013).

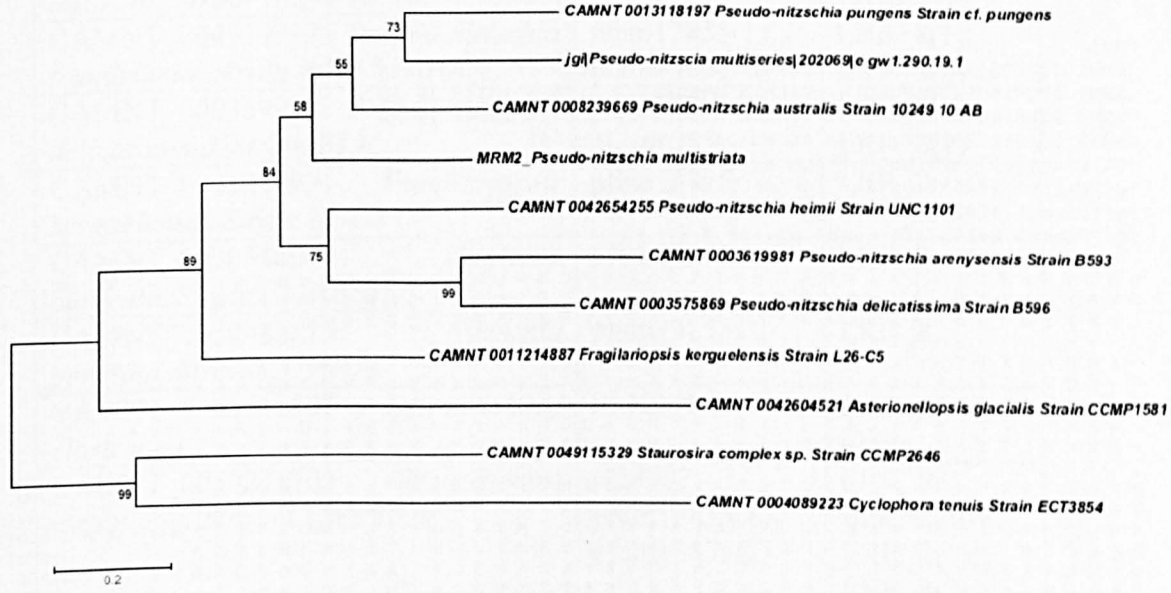


Figure 2.15: : Molecular Phylogenetic analysis by Maximum Likelihood method of gene *MRM2*. The evolutionary history was inferred by using the Maximum Likelihood method based on the

Whelan And Goldman model (Whelan and Goldman 2001). The tree with the highest log likelihood (-4498.7455) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using a JTT model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.4666)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 11 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 241 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.* 2013).

The phylogenetic clustering among the diatom species was consistent with the phylogenetic relationships based on SSU rDNA (Kooistra *et al.* 2003). All the MT-biased genes appear to have a single homolog in the other species except for *MRP1*, for which two paralogs exist in *Pseudo-nitzschia multiseries*, *Pseudo-nitzschia australis* and *Fragilariopsis cylindrus;* and for *MRP2*, for which two paralogs exist in *Nitzschia punctata.*

The sequence alignments produced by ClustalW are reported in APPENDIX C. Moreover they were manually curated before performing phylogenetic analysis.

All the homologs identified for *MRP1*, with the exception of protein CAMNT 0008148059 of *Pseudo-nitzschia australis*, presented the cleavage site of signal peptide (VSA-DY or SAA-EY) at the beginning of sequence and the conserved motif EH—WEKLFC at position 150 as illustrated in the two alignment fragments (Fig. 2.16).

```
MRP1
CAMNT 0013081181Pseudo-nitzschia pungens Strain cf. cingulata
CAMNT 0008160915Pseudo-nitzschia australis Strain 10249 10 AB
CAMNT 0008148059Pseudo-nitzschia australis Strain 10249 10 AB
CAMNT 0020483251Nitzschia punctata Strain CCMP561
jgi|Fracy1|271829|estExt fgenesh2 kg.C 300013
jgi|Fracy1|268858|estExt fgenesh2 kg.C 50302
jgi|Psemu1|4292|gm1.4292 g
jgi|Psemu1|291017|fgeneshl pg.603 * 5

MMTFNFSTVVLALVAAT-SFVSADYVCENQAFFKLDTKKKPSK----K [50]
MMMKIFATALALVAAS--PVVSAIYHCESETTFQLEDSVKPSN----A [50]
-MMKFATLALALVAAS-TPLVSAEYTCHGETYFQFEDGS-ISKP-SPE [50]
----------MIVSA--HPLK---------------------------Q [50]
--MKLSFVFAVLSAVVAPVANANLYKCETSAVFDVEDDATKPQVPSKE [50]
--MQFSTIALLLAALI--APSAAEYVCHSDASFNSDVDTMPSA----A [50]
--MKFSTIALLLTTLI--APSAAEYVCHSDASFNSDVDSMPSP----A [50]
-MMKFCTFALALIAT--FTIVSADYQCESSTTFHWADGSKPSKP-SKA [50]
-MMKFLTTALALVAASPIASVSANYQCESDTSFVMADAVAPSK----E [50]

MALTTSSEHRLWEKLFCQKARTNKDFKTISGCSIVLSDCHNENG-----[200]
VALATSSEHKAWEKLFCERAHQMKEFATMEKCAIVLKDCGEANDG-----[200]
AALTSSMEHKVWEKLFCEYASAHEEFDTMTSCSIVLSNCHKESASEHESE[200]
MALTTSIEHKNWEKLFCEKASNLKKFSSMTDCSIVLSNCHKESA-----[200]
DFVAASSEHKEWEKLFCAGIKKNPEFASAKGCAIALTNCQDDGEANNDVD[200]
VALLSSKEHVAWEKLFCAKGRANSEFTSMTDCKIDLSNCHDDDE---DNV[200]
IALSSSKEHLSWEKLFCAKGSANAEFSSMTDCKIDLSNCHDDEV---DYV[200]
LALNTSAEHSSWEKLFCEKVHSRKSFSTLTGCAIHLNNCETEPP-----[200]
LALNTAAEHSSWEKLFCEKVHKLEEFASMTGCAIHLTNCETTTE-----[200]
```

Figure 2.16: Fragments of MRP1 alignment presenting in red the signal peptide (VSA-DY or SAA-EY) and the conserved motif EH—WEKLFC.

Although the results appear interesting no clarifying information were gained to better characterize *MRP1*.

Concerning *MRP2*, it was identified that all the homologs proteins had the LLR_8 domain conserved (Table 2.12), except for *Pseudo-nitzschia australis*, *Cylindrotheca closterium*, *Pseudo-nitzschia multiseries* and *Fragilariopsis cylindrus*, which showed an incomplete domain of leucin rich repeats. In addition, *Pseudo-nitzschia pungens* and *Thalassiosira weissflogii* protein had a Leucine-rich repeats (LRRs), ribonuclease inhibitor (RI) conserved domain.

Table 2.12: Batch CD search tool of NCBI to analyse conserved domain of *MRP2* and its homolog proteins alignment.

Sequence	Accession	Short name
MRP2	pfam13855	LRR_8
CAMNT_0013082077 *Pseudo-nitzschia pungens, Strain cf. cingulate*	cl23707, pfam13855	*Incomplete* LRR_RI superfamily, LRR_8
CAMNT_0008163717 *Pseudo-nitzschia australis, Strain 10249 10 AB*	PLN00113	*Incomplete* LRRs
CAMNT_0047397535 *Pseudo-nitzschia fraudulenta, Strain WWA7*	pfam13855(1), pfam13855(2)	LRR_8(1), LRR_8(2)
CAMNT_0003599921 *Pseudo-nitzschia delicatissima, Strain B596*	pfam13855(1), pfam13855(2)	LRR_8(1), LRR_8(2)
CAMNT_0041254931 *Fragilariopsis kerguelensis, Strain L2-C3*	pfam13855	LRR_8
CAMNT_0020546941 *Nitzschia punctata, Strain CCMP561*	pfam13855	LRR_8
CAMNT_0020483817 *Nitzschia punctata, Strain CCMP561*	pfam13855	LRR_8
CAMNT_0000511607 *Cylindrotheca closterium*	PLN00113	*Incomplete* LRRs
CAMNT_0050271705 *Thalassiosira weissflogii, Strain CCMP1010*	cl23707, pfam13855(1), pfam13855(2)	LRR_RI superfamily, LRR_8(1), LRR_8(2)
jgi|*Pseudo-nitzschia multiseries*|98760|gw1.572.7.1 MANUALLY CURATED_stop codon added	PLN00113	*Incomplete* LRRs

jgi|*Fragilariopsis cylindrus*|257266|fgenesh2_pg.89_#_34	PLN00113	*Incomplete* LRRs

As reported before, *MRP3* is an uncharacterised protein. The Batch CD search tool gave the same result for all its homologues proteins except for *Fragilariopsis cylindrus* jgi|Fracy1|272356|estExt_fgenesh2_kg.C_440048 protein that, although very short (99 a.a.), found match as incomplete hit of PRK11239, member of the superfamily cl01209 name DUF480 of unknown function and of bacterial origin.

Batch CD search for the protein of *MRM1* and its homologs identified for all of them the same conserved domain of Heat Shock Factor DNA-binding (Table 2.13).

Table 2.13: Batch CD search tool of NCBI to analyse conserved domain of *MRM1* and its homolog proteins alignment.

Sequence	Accession	Short name	Superfamily
MRM1	pfam00447	HSF_DNA-bind	cl12113
CAMNT_0013145689 *Pseudo-nitzschia pungens, Strain cf. pungens*	pfam00447	HSF_DNA-bind	cl12113
CAMNT_0042638377 *Pseudo-nitzschia heimii, Strain UNC1101*	pfam00447	HSF_DNA-bind	cl12113
CAMNT_0011396591 *Fragilariopsis kerguelensis, Strain L26-C5*	pfam00447	HSF_DNA-bind	cl12113
CAMNT_0000335841 *Nitzschia sp.*	pfam00447	HSF_DNA-bind	cl12113
CAMNT_0042665851 Skeletonema marinoi, Strain UNC1201	pfam00447	HSF_DNA-bind	cl12113
jgi| *Pseudo-nitzschia multiseries*|282196|fgenesh1_pg.4_#_2	pfam00447	HSF_DNA-bind	cl12113
jgi| *Fragilariopsis cylindrus*|250502|fgenesh2_pg.31_#_101	pfam00447	HSF_DNA-bind	cl12113

Batch CD search analysis of *MRM2*, whose protein annotation was corrected as a probable Leucin rich repeat (LRR)-containing protein, showed for some of the homologs proteins an incomplete domain of leucin rich repeats. The homologous proteins of *Pseudo-nitzschia heimii, Fragilariopsis kerguelensis, Staurosira complex sp.* and *Cyclophora tenuis* had the LLR_8 domain conserved; while only *Staurosira complex sp.* and *Asterionellopsis*

glacialis presented a Leucine-rich repeats (LRRs), ribonuclease inhibitor (RI) conserved domain (Table 2.14).

Table 2.14: Batch CD search tool of NCBI to analyse conserved domain of *MRM2* and its homolog proteins alignment.

Sequence	Accession	Short name
MRM2	PLN00113	*Incomplete* PLN00113
CAMNT_0013118197 *Pseudo-nitzschia pungens, Strain cf. pungens*	PLN00113	*Incomplete* PLN00113
CAMNT_0042654255 *Pseudo-nitzschia heimii, Strain UNC1101*	pfam13855	LRR_8
CAMNT_0003619981 *Pseudo-nitzschia arenysensis, Strain B593*	PLN00113	*Incomplete* PLN00113
CAMNT_0003575869 *Pseudo-nitzschia delicatissima, Strain B596*	PLN00113	*Incomplete* PLN00113
CAMNT_0011214887 *Fragilariopsis kerguelensis, Strain L26-C5*	pfam13855	LRR_8
CAMNT_0008239669 *Pseudo-nitzschia australis, Strain 10249 10 AB*	PLN00113	*Incomplete* PLN00113
CAMNT_0049115329 *Staurosira complex sp., Strain CCMP2646*	cl23707, pfam13855	LRR_RI superfamily, LRR_8
CAMNT_0004089223 *Cyclophora tenuis, Strain ECT3854*	pfam13855	LRR_8
CAMNT_0042604521 *Asterionellopsis glacialis, Strain CCMP1581*	cl23707	LRR_RI superfamily
jgi|*Pseudo-nitzschia multiseries*| 202069|e_gw1.290.19.1	PLN00113	*Incomplete* PLN00113

2.3.8 Selective pressure acting on *P. multistriata* MT-biased genes

The set of 91 genes, resulted differentially expressed, was searched in the set of 6066 genes produced by the Ka/Ks calculation. Of the 91 genes, only 61 resulted to have a correct gene model prediction and, of these, only 23 were found to have an orthologue with *P. multiseries* (Table 2.15) for which a Ka/Ks value was calculated. Only two genes were under positive selection and one of them was *MRP1* (PSNMU-V1.4_AUG-EV-PASAV3_0024820.1) with a Ka/Ks value >1. The gene PSNMU-V1.4_AUG-EV-PASAV3_0003630.1, corresponding to the transcript "comp31789_c0_seq1.1", resulted to

be under positive selection with a Ka/Ks value >1. Although this gene was in the list of 91 putative MT-biased genes, the validation analyses showed that it was not differentially expressed nor related to mating type.

MRP2 (PSNMU-V1.4_AUG-EV-PASAV3_0122240.1) instead presented a Ka/Ks value <1, that does not mean positive selection is not occurring. It can happen that the mutations are neutral or disadvantageous, or some of the mutations are advantageous and some disadvantageous resulting in a Ka/Ks ratio in the range 0 to 1.

The remaining three MT-biased genes *MRP3*, *MRM1* and *MRM2*, did not show an orthologous gene in *P. multiseries*. This discrepancy with the results shown in figure 2.10 is probably due to the fact that Ka/Ks calculations are made using nucleotide sequences (CDS) while the TBLASTN searches were made using protein sequences.

Table 2.15: Ka/Ks ratio of *P. multistriata* MT-biased genes. In the table are reported *P. multistriata* transcript name, *P. multiseries* transcript name, Ka value, Ks value, Ka/Ks value, P-Value of the Fisher test (null hypothesis: Ka/Ks = 1), FDR: p-value corrected for multiple testing (Benjamini-Hochberg FDR), Description: description of the transcript in *P. multistriata*.

P. multistriata	*P. multiseries*	Ka	Ks	Ka/Ks	P.Value Fisher.	FDR	Description
PSNMU-V1.4_AUG-EV-PASAV3_0003630.1	jgi|Psemu1|300843|fgenesh1_kg.13_#_31_#_6665_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS	0,44	0,19	**2,28**	7,14E-02	7,27E-02	
PSNMU-V1.4_AUG-EV-PASAV3_0020420.1	jgi|Psemu1|167081|gw1.43.201.1	0,09	3,61	0,03	0,00E+00	0,00E+00	
PSNMU-V1.4_AUG-EV-PASAV3_0022580.1	jgi|Psemu1|321664|estExt_fgenesh1_pm.C_24880001	0,09	1,81	0,05	0,00E+00	0,00E+00	
PSNMU-V1.4_AUG-EV-PASAV3_0024820.1	jgi|Psemu1|182540|e_gw1.24.83.1	0,17	0,08	**2,13**	2,74E-04	2,83E-04	
PSNMU-V1.4_AUG-EV-PASAV3_0024	jgi|Psemu1|150676|gw1.24.146.1	0,19	3,23	0,06	0,00E+00	0,00E+00	

880.1							
PSNMU-V1.4_AUG-EV-PASAV3_0036200.1	jgi|Psemu1|296887|fgenesh1_pm.174_#_3	0,1	2,27	0,04	2,61E-143	6,82E-143	Delta(12) fatty acid desaturase fat-2
PSNMU-V1.4_AUG-EV-PASAV3_0045400.1	jgi|Psemu1|692005|estExt_Genemark1.C_6170025	0,1	1,45	0,07	2,16E-103	4,06E-103	
PSNMU-V1.4_AUG-EV-PASAV3_0045410.1	jgi|Psemu1|213192|e_gw1.617.23.1	0,11	1,47	0,08	1,27E-94	2,21E-94	
PSNMU-V1.4_AUG-EV-PASAV3_0059450.1	jgi|Psemu1|706666|estExt_Genemark1.C_17980002	0,07	2,69	0,03	4,58E-118	9,77E-118	
PSNMU-V1.4_AUG-EV-PASAV3_0062100.1	jgi|Psemu1|183037|e_gw1.27.151.1	0,06	0,92	0,07	1,59E-57	2,07E-57	
PSNMU-V1.4_AUG-EV-PASAV3_0064960.1	jgi|Psemu1|325712|estExt_fgenesh1_pg.C_2310017	0,03	0,69	0,05	2,82E-63	3,81E-63	
PSNMU-V1.4_AUG-EV-PASAV3_0072310.3	jgi|Psemu1|708444|estExt_Genemark1.C_23640003	0,24	2,57	0,09	0,00E+00	0,00E+00	
PSNMU-V1.4_AUG-EV-PASAV3_0090370.1	jgi|Psemu1|253267|estExt_Genewise1Plus.C_580075	0,18	1,17	0,16	1,43E-32	1,60E-32	
PSNMU-V1.4_AUG-EV-PASAV3_0091290.1	jgi|Psemu1|171808|gw1.793.32.1	0,11	1,98	0,05	1,53E-57	1,99E-57	
PSNMU-V1.4_AUG-EV-PASAV3_0103000.1	jgi|Psemu1|15307|gm1.14964_g	0,23	2,37	0,1	0,00E+00	0,00E+00	
PSNMU-V1.4_AUG-EV-PASAV3_0105520.1	jgi|Psemu1|261038|estExt_Genewise1Plus.C_4850020	0,07	2,16	0,03	7,71E-122	1,70E-121	Elongation of very long chain fatty acids protein 2
PSNMU-	jgi|Psemu1|2838	0,17	1,64	0,1	3,53E-	8,97E-	

V1.4_AUG-EV-PASAV3_0108790.1	43|fgenesh1_pg.28_#_30				140	140	
PSNMU-V1.4_AUG-EV-PASAV3_0113670.1	jgi|Psemu1|262333|estExt_Genewise1Plus.C_6630010	0,04	1,1	0,03	1,53E-115	3,18E-115	Delta(12) fatty acid desaturase
PSNMU-V1.4_AUG-EV-PASAV3_0113740.1	jgi|Psemu1|291106|fgenesh1_pg.555_#_1	0,19	1,27	0,15	0,00E+00	0,00E+00	Uncharacterized helicase C15C4.05
PSNMU-V1.4_AUG-EV-PASAV3_0116760.1	jgi|Psemu1|282609|fgenesh1_pg.5_#_7	0,09	1,33	0,07	2,97E-89	4,92E-89	Pumilio homolog 2
PSNMU-V1.4_AUG-EV-PASAV3_0117860.1	jgi|Psemu1|325408|estExt_fgenesh1_pg.C_2000012	0,09	1	0,09	5,97E-64	8,12E-64	GDT1-like protein 2 chloroplastic
PSNMU-V1.4_AUG-EV-PASAV3_0121970.1	jgi|Psemu1|312609|fgenesh1_kg.906_#_6_#_5595_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS	0,05	0,71	0,08	3,33E-38	3,84E-38	
PSNMU-V1.4_AUG-EV-PASAV3_0122240.1	jgi|Psemu1|99103|gw1.572.7.1	0,21	0,43	0,49	1,08E-06	1,12E-06	LRR receptor-like serine/threonine-protein kinase GSO1

2.4 Discussion

2.4.1 Sex (MT)-biased genes

The work illustrated in this chapter is the first differential gene expression study between opposite mating types in diatoms and the second attempt to identify the MT-locus, after the publication of the AFLP-based linkage map approach carried out on the benthic diatom *Seminavis robusta* (Vanstechelman *et al.* 2013). The analysis conducted on *Pseudo-nitzschia multistriata* was designed to search for the constitutive differences in gene expression between opposite mating types, meaning differences observed in the vegetative cells, both above (>SST) and below (<SST) the sexualisation size threshold. The purpose was to detect the MT-determining gene/s and the MT-biased ones.

The first aim has not been achieved with the analyses presented here. If, as expected for a genetic determination of sex, one of the two mating types is heterogametic, one would expect the absence of a given allele in the other mating type. None of the MT-biased genes identified in this study was absent in the genome of the strains for which expression was not detected (*MRM1* was not expressed in MT+ strains but it could be found in their genome), nor was there any specific polymorphic pattern which could be compatible with the gene being a primary MT-determining gene (see Chapter 4). Possible explanation for the failure in identifying the primary MT-determining gene/s could be that this gene is not expressed in the conditions considered in this study, i.e. in strains growing in monoculture but rather it is expressed in a mating type-specific manner during a particular stage, e.g., in a very short time window of the life cycle, such as in concomitance with the switch between >SST and <SST. In this case, the sex-determining gene could trigger a cascade of events that would lead to stable expression/repression of the genes that were expressed by one of the two MT.

The analysis presented in this chapter proved that a set of male (MT+)-biased and female (MT-)-biased genes are expressed in sexually mature (<SST) cultures. It has been already reported that, among multicellular organisms, sex-biased gene expression becomes most pronounced after sexual differentiation (Ellegren and Parsch 2007). Moreover, sex-biased gene expression appears to be dynamic throughout development in a number of species (Ingleby *et al.* 2014). As the sexes differentiate, the expression of sex-biased genes increases, since sexually antagonistic selection is likely to be stronger when distinct male and female traits are specified and produced (Ingleby et al., 2014).

The genomes of male (MT+) and female (MT-) individuals differ by only a few genes located on sex chromosomes or sex regions, meaning that their sexual dimorphisms results from the differential expression of genes present in both sexes. Sex (MT)-biased genes include those that are expressed exclusively in one sex, so called sex-specific, as well as those that are expressed in both sexes but at a higher level in one, so called sex-enriched. Depending on which sex shows higher expression, sex-biased genes can be further separated into male-biased and female-biased genes (Ellegren and Parsch 2007). Given the definition of sex (MT)-biased genes by Ellegren and Parsch (2007) the results of the analyses illustrated in this Chapter (Figs 2.2-2.6) showed that, among the male-biased genes, two (*MRP1* and *MRP3*) were found to be MT+ specific and the third one (*MRP2*) MT+ enriched. Among the female-biased genes, one (*MRM1*) was MT- specific and one (*MRM2*) MT- enriched.

Sex-related differences in gene expression are reported across a wide range of taxa: insects, nematodes, birds, mammals and also algae (Martins *et al.* 2013, Patil 2014, Lipinska *et al.* 2015). In the brown alga *Fucus vesiculosus*, a pattern of greater expression of male-biased genes was shown by comparing male and female sexual tissues where 92 and 28 over-expressed genes, respectively, were identified (Martins *et al.* 2013). In the haploid stage of the brown alga *Ectocarpus* gene expression patterns during sexual differentiation were

measured in the juvenile immature gametophytes and at sexual maturity. Male-biased genes were more numerous than female-biased ones at both developmental stages. However, the overall number of genes differentially transcribed between males and females was higher during the immature gametophyte stage than at gametophyte fertility. In total, fewer than 12% of *Ectocarpus* genes exhibited sex-biased expression (Lipinska *et al.* 2015)

Patil (2014) performed a similar comparative analysis on *Pseudo-nitzschia multistriata*, where the overall number of genes differentially transcribed between males and females was higher among the sexualised samples (two cultures of different mating type growing separated by a filter, but sharing the chemical contact through the culture medium) than in the vegetative ones (monoculture of a single mating type). Moreover, the comparison between MT+ sexualised cell type against MT- sexualised yielded 36 transcripts uniquely up-regulated in sexualised MT+ and 182 transcripts uniquely up-regulated in sexualised MT-. This result is apparently opposite to what observed in *Fucus* and *Ectocarpus*, where male-biased genes were more numerous. However it has to be remembered that the definition of male (MT+) and female (MT-) for pennate diatoms is not the same as for brown algae. In *P. multistriata*, the definition of MT+ (male) and MT- (female) strains is arbitrary. The strain that holds the auxospores is defined as MT- (female). It thus might well be that the mating types of *P. multistriata* have an opposite assignment.

Ingleby *et al.* (2014) summarized some of the studies focused on sex-biased genes on a wide range of taxa (mammals, fishes, birds, amphibians, nematodes, platyhelminthes, crustaceans, molluscs and insects) showing that there is no general trend for how much the transcriptome is sex-biased and wide variations can be observed even between the same species. For example, only about 2% of the transcriptome of the marine snail *Littorina saxatilis* was found to be sex-biased whereas in *Drosophila melanogaster* sex-bias covers 90% of the transcripts. The author listed a number of potential explanations for this variation, including tissue specificity of gene expression, developmental stage,

intraspecific genetic and environmental variation and the experimental design and analytical techniques specific to each study (Ingleby *et al.* 2014).

Lipinska *et al.* (2015) explained the low percentage of *Ectocarpus* in sex-bias as consistent with the low level of sexual dimorphism in this species. Those conclusions could perhaps explain the low percentage of MT-biased genes detected and validated in *P. multistriata* during the vegetative stage, showing that few genes are responsible for the determination of the mating type phenotype. Also *P. multistriata* shows very low levels of sexual dimorphism; the two mating types are not morphologically, nor functionally distinct except for the different behaviour during fertilization, when the MT+ gametes glide towards the MT- ones (Scalco *et al.* 2015) and for the auxospore development that occurs on the MT- gametangium. Not only one developmental stage was included in the RNA-Seq. In fact, the transcriptomic data included sequences from cells <SST and >SST, which were essential to test the expression of the five MT-biased genes in the large immature (>SST) cells. The results of this latter analysis will be illustrated and discussed in Chapter 3.

2.4.2 Methodological considerations

Gene expression can be variable among individuals of any species, and the comparison of multiple strains in this study was challenging because natural fluctuations in the basal levels of expression of a given gene could lead to ambiguous results. A special care has been taken to make sure that enough samples were considered, that the starting material for qPCR was of high quality and that no major technical bias were present. When designing a qRT-PCR validation experiment, three important aspects should be considered.

First: a careful assessment of the number of biological samples needed to draw meaningful and statistically significant results is needed (Derveaux *et al.* 2010). During my project, I experienced that qRT-PCR validations based on at least four samples (for each MT) were able to discriminate those transcripts differentially expressed in a MT-specific manner

from others, equally differential expressed, but in a strain-specific way. High numbers of biological replicas are thus essential not only to produce a transcriptome but also to validate it, so to lower the expression variability due to strain specificity.

Second: to choose between a sample maximization or a gene maximization strategy (Derveaux *et al.* 2010). The choice is related to the biological question. However, in a relative quantification study, the experimenter is usually interested in comparing the expression level of a gene between different samples. Therefore, the sample maximization method is highly recommended, to reduce the run-to-run variation between the samples; this is what I did.

Third: the RNA samples quality is crucial as it highly impacts on the results. It is important to perform a quality RNA check analyzing its quality score (RIN or RQI) or to use PCR-based tests to determine mRNA integrity (Derveaux *et al.* 2010). Moreover, it is also recommended to run qPCR analysis to check for absence of DNA after a proper DNase treatment.

The first set of samples was not homogenous in terms of RNA extraction dates. Nevertheless, the results obtained in the two validations sets were consistent, although the first series was less selective than the second (data not shown). This shows that the quality of the extracted RNA was good, notwithstanding the difference in storage time. I also tested that a single PCR run amplifying for 1Kb fragment of a reference gene containing an intronic region was enough to check for RNA integrity and DNA contamination.

2.4.3 Characterization of *Pseudo-nitzschia multistriata* MT-biased genes

The percentage of annotated transcripts of *P. multistriata* within the 91 differentially expressed MT+/MT- genes was only 17% compared to 67% of the total transcriptome. This suggests that unique molecular mechanisms regulate the mating process.

The differential expression analysis represents a powerful resource for identifying candidate diatom-specific genes involved in processes of major ecological relevance and for gene annotation in diatoms and related genomes (Bowler *et al.*, 2010). Differential expression studies related to sex (MT)-biased genes were never conducted on other diatom species, whereas few examples are available for brown algae that cluster together with diatoms in the Stramenopile clade. The study of the functional role detected for the sex (MT)-biased genes during sexualisation of *E. siliculosus*, *F. vesiculosus* and *P. multistriata* (Martins *et al.* 2013, Patil 2014, Lipinska *et al.* 2015) could possibly clarify the function of the five MT-biased genes in the latter. However, it is necessary to remember that the characters based on which the designation of sexes is based is different between brown algae and diatoms, and, in the latter, is arbitrary. The functional analysis of sex-biased genes through gene ontology (GO) enrichment in the male (MT+) biased genes in mature gametophytes of *E. siliculosus* resulted in enrichment of specific GO categories for "microtubule" and "calcium binding-related" processes (Lipinska *et al.* 2013, Lipinska *et al.* 2015). In *F. vesiculosus* male sexual tissue, signalling-related genes and genes related to flagella localization and functions were overrepresented (Martins *et al.* 2013). Oxydoreductase, monoxygenase, serine type peptidase, inositol dephosphorylation, intracellular signal transduction and iron binding processes were overrepresented in MT+ sexualised samples of *P. multistriata* (Patil 2014). The signalling processes resulted to be a common denominator of all male (MT+)-biased genes for the three species. On the other side, the set of female (MT-) biased genes in juvenile gametophytes of *E. siliculosus* were enriched of specific GO categories for "photosynthesis" (Lipinska *et al.* 2015) and in female sexual tissue of *F. vesiculosus* they were enriched of carbohydrate-modifying enzymes (Martins *et al.* 2013). dUTP diphosphatase activity, DNA specific nucleotide binding transcription activity, ubiquitination and proteolysis activities were, instead, overrepresented in MT- sexualised strains of *P. multistriata,* suggesting that MT- strains may be involved in protein internalization to initiate a signalling process and degradation

of the potential pheromone or of the cell surface receptors for pheromone perception (Patil 2014). Indeed, the two MT- biased genes *MRM1* and *MRM2,* encoding for proteins with a HSF-type DNA-binding domain and a LRR receptor-like protein, respectively, may have the following function: *MRM1* could activate a transcriptional cascade to initiate the signalling process once that *MRM2,* working as a receptor, has internalized the chemical cue.

It is known that HSF are the major regulators of heat shock protein transcription in eukaryotes. Heat shock factors trigger the expression of genes encoding heat shock proteins (HSP) that function as molecular chaperones. Nevertheless, their function is not only critical to overcome the proteotoxic effects of thermal stress, but also for performing crucial roles during gametogenesis and development in standard conditions. HSFs regulate very specific sets of heat shock genes, but also many other genes encoding growth factors or involved in cytoskeletal dynamics (Abane and Mezger 2010). It was found that the presence or absence of HSP influence various aspects of sexual reproduction in many species. In humans, they can even act as antigens of numerous microbial pathogens that can cause infertility (Neuer *et al.* 2000), while in *Drosophila* they play a fundamental role in oogenesis (Marin and Tanguay 1996). In *Caenorhabditis elegans* the transcription factor HSF1 was found to influence aging; reducing HSF1 activity accelerates tissue aging while its overexpression extends life span (Hsu *et al.* 2003).

LRRs occur in proteins ranging from prokaryotes to eukaryotes, and appear to provide a structural framework for the formation of protein-protein interactions (Petersen *et al.* 2011, Gissendanner and Kelley 2013). Proteins containing LRRs include tyrosine kinase receptors, cell-adhesion molecules, virulence factors, and extracellular matrix-binding glycoproteins, which are involved in a variety of biological processes, including signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, apoptosis, and the immune response (Ashworth *et al.* 2013, Letunic *et*

al. 2015, Schulze *et al.* 2015). A LRR receptor-like conserved domain with a transmembrane region at the N-terminus of the protein was present in the MT+ biased gene *MRP2* in *P. multistriata*, leading to the hypothesis that both mating types are involved in perception of external cues and signal transduction. In diatoms LRR is one of the more represented protein families. The first in-depth analysis of LRR proteins encoded by *Phaeodactylum tricornutum* was performed by Shulze *et al.* (2015). The Authors were able to identify several transmembrane LRR-proteins, which are likely to function as receptor-like molecules and several secreted LRR proteins likely to function as adhesion or binding proteins as part of the extracellular matrix. However, their structures were quite different from mammalian or plant-like receptors leading to the conclusion that signal recognition pathways are substantially different in diatoms.

A receptor-like kinase (CpRLK1) was found to be a candidate key factor involved in fertilization in the Charophycean alga *Closterium peracerosum-strigosum-littorale* complex through a microarray expression analysis (Sekimoto *et al.* 2006). The knockdown of CpRLK1 in MT+ showed reduced competence for sexual reproduction after pairing with MT– cells. The knockdown cells were unable to release the naked gamete and formed an abnormally enlarged conjugation papilla, thus impairing conjugation and zygote formation (Hirano *et al.* 2015). The Authors suggested that the CpRLK1 protein is an ancient cell wall sensor that now functions to regulate osmotic pressure in the cell to allow proper gamete release. Many studies on the role of receptor-like protein kinase during sexual reproduction have been produced on plants. The processes of pollen tube attraction, growth arrest, bursts and release of the sperm cells are controlled by the female gametophyte *via* the FERONIA receptor-like protein kinase (FER- RLK) in *Torenia fournieri* (Escobar-Restrepo *et al.* 2007). In *Arabidopsis,* two close homologs of FER-RLK are expressed in the pollen tube and enable it to break at the appropriate time to deliver sperm cells (Boisson-Dernier *et al.* 2009, Miyazaki *et al.* 2009).

The MT+ specific gene *MRP1* of *P. multistriata* resulted not annotated, as most of the transcripts in the list of the up-regulated genes of the MT+ samples. Commonly genes expressed in sperm or in the male germ line showed a significant excess of 'orphans', i.e. genes that did not give a significant BLAST match between the species. This can be possibly explained by the higher evolution rate in male sex-biased genes of diploid systems (Ellegren and Parsch 2007, Ingleby *et al.* 2014). *MRP1* presented a signal peptide at the N-terminus of the protein and was predicted to have an extracellular localization. Signal peptides are generally short sequence peptides present at the N-terminus of the majority of newly synthesized proteins that are destined towards the secretory pathway. Signal peptides are found in proteins that are targeted to the endoplasmic reticulum and eventually destined to be either secreted, retained in the lumen of the endoplasmic reticulum, of the lysosome or of any other organelle along the secretory pathway or to be I single-pass membrane proteins. The signal sequence is usually removed in the mature protein. Peptides are excellent signals in marine systems given their high solubility, short half-lives due to rapid consumption by bacteria, and correspondingly high signal to noise ratios (Rittschof and Cohen 2004). Many examples of crustacean peptide and peptide-like pheromones, and the processes which the pheromones are involved in, have been reviewed. Peptides have been shown to attract consumers, both in laboratory experimental conditions or in various field observations, suggesting that they are commonly used to find foods or other resources (Rittschof and Cohen 2004, Hay 2009). For example, the waterborne pheromones, used in barnacle settlement, have been the first to be described as a heterogeneous group of peptides between 1000 and 10,000 Da. with other peptides <500 Da. The hypothesis fostered was that all of the smaller active molecules were serine protease degradation products (Rittschof and Cohen 2004). Again this hypothesis of serine protease degradation was used to explain the mixture of di/tripeptides in the crustacean larval release of the pumping pheromone (Rittschof and Cohen 2004) . The modified amino sugars released from fish mucus fragments and ctenophore predators (10–30 kDa) turned out to be active

cue molecules (<10 kDa) when hydrolysed with either bacterial heparinase or chondroitinase, inducing the avoidance response of crab's larvae (Rittschof and Cohen 2004). Many peptides pheromone are involved also in sexual reproduction processes of different marine organisms. In the marine ragworm *Nereis succinea*, a tetra-peptide cysteinyl–glutathione (CSSG) was detected as mate recognition and gamete release pheromone during reproduction (Hardege *et al.* 2004). The pheromone induced not only gamete release in males but, already in low doses, it also significantly increased male swimming activity (Hardege *et al.* 2004).

For what concern unicellular microalgae, a mating-type specific gene, AT4-3, identified in the dinoflagellate *Alexandrium tamarense* was found to be differentially expressed in one of the mating types (Kobiyama *et al.* 2007). The predicted amino acid sequence of AT4-3 had a presumptive N-terminal signal peptide for extracellular secretion, but still the gene has no annotation and no clear functional role. Three of the sexually induced genes, Sig1, Sig2, and Sig3 of *Thalassiosira oceanica* encode for three polypeptides, each possessing a putative signal sequence characterized by a stretch of 12 to 14 hydrophobic amino acids preceded at the N-terminus by one or two basic residues, and cysteine-rich epithelial growth factor (EGF)-like repeats (Armbrust 1999). It was found a striking similarity between the SIG polypeptides and the extracellular matrix components, commonly involved in cell-cell interactions, suggesting that the SIG polypeptides may play a role in sperm-egg recognition (Armbrust 1999). It has been recently found that these domains encode for components of stramenopile mastigonemes (Honda *et al.* 2007). The MT- attracting pheromone of *S. robusta* (Chapter 1.3.2 and 3.4.1) was identified as a cyclic dipeptide derived from two proline moieties (Gillard *et al.* 2013, Frenkel *et al.* 2014).

Therefore, it can be hypothesized that *MRP1* can possibly act as a pheromone towards the MT- cells, in which an up-regulated Cathepsin D (pepsin-like aspartate protease) - that has been shown to cleave proteins in the extracellular matrix (Handley *et al.* 2001) - stands out. This function is very interesting as *MRP1* could encode for a candidate MT+ pheromone

and cathepsin D could be the potential extracellular protease that cleaves the pheromone secreted by the MT+ strain.

It is clear from the reported examples that: i) pheromone chemistry in algae is highly diverse, produced by several different pathways that include ribosomal protein production, fatty acid catabolism, and terpenoid pathways; ii) the molecular weight range extends from small hydrocarbons to large protein complexes. Concluding, it can be hypothesized that the five MT-biased genes of *Pseudo-nitzschia multistriata* may be involved in the sex determination system as downstream regulated genes.

2.4.4 Conservation of *Pseudo-nitzschia multistriata* MT-biased genes

The results of the analyses of conservation showed that the five MT-biased genes of *P. multistriata* were restricted only to diatom species. These results were confirmed also by the findings of Basu *et al.*, (under revision) who carried out an extensive phylogenomic study based on bacterial (1116 species) and archaeal (121 species) proteomes and eukaryotic proteomes (from 50 sequenced genomes) broadly representing the tree of life. The authors conducted a phylogenetic clustering analysis, generating ~240,000 clusters of putative homologous proteins, out of which 8113 contained 9122 *P. multistriata* proteins. Searching in the 8113 clusters of putative homologous proteins, I could not find any cluster containing either *MRM1* or *MRM2* while these two gene were found in the list of orphan genes. This discrepancy with my results, which showed that the five MT-biased genes were all conserved to a certain degree, can be explained by the stringent cutoff selected by the authors (>50% sequence identity) for the homologs identification.

The absence of conservation of gene/s related to MT in the species for which only the transcriptome was available is not a strong evidence as their absence in the genome. Since a transcriptome represents the transcriptional activity of a cell at the moment in which the sample was collected, the absence of a transcript could depend on different reasons, e.g.,

the gene encoding for the transcript was switched off, or the transcript was expressed at a very low levels.

Another important factor to be considered is that we do not know the mating type and the cell size of the pennate diatom species included in the MMETSP dataset. If the strains were above their species-specific sexualisation size threshold, the MT-biased gene was presumably not expressed. For the only strain for which we know the MT, the *Pseudo-nitzschia arenysensis* strain B593 (transcriptome MMETSP0329) which was a MT- strain (M. Ferrante, personal communication), the absence of MT+ biased genes is consistent with my findings. Another factor to be considered is that two of the five genes, i.e. *MRP2* and *MRM2*, were MT-enriched and not MT-specific. It might therefore be that they were still detected in an MT in which we would have predicted absence because they were not completely off but rather were present with a very low level of expression.

Four out of five MT-biased genes presented homologs in both *Fragilariopsis cylindrus* and *Pseudo-nitzschia multiseries,* for which the analyses were conducted on the genome. Both MT+ biased and MT- biased genes were detected, but without their expression levels no correlation to mating type can be made. *Pseudo-nitzschia multiseries* is congeneric to *P. multistriata* while the genus *Fragilariopsis* is phylogenetically very close to *Pseudo-nitzschia* (Lundholm *et al.* 2002, Kooistra *et al.* 2003) and this can explain the conservation of the four MT-biased genes. The life cycle of *F. cylindrus* is not known; however a heterothallic life cycle has been reported for *F. kerguelensis* (Fuchs *et al.* 2013) that actually shows conservation of two MT+ biased genes in the transcriptome of Strain L2-C3 and of the MT- biased one in the transcriptome of Strain L26-C5. Although information on the MTs of *F. kerguelensis* strains is not available, it could be hypothesized that the two strains reported in Fig. 2.9 had opposite mating type.

The absence of conservation for the majority of the MT-biased genes among the phylogenetically distantly related species such as *Seminavis robusta, Skeletonema marinoi, Phaeodactylum tricornutum, Thalassiosira pseudonana* and *Ectocarpus siliculosus* and

absence of all these genes in all the eukaryotic species included in the MMETSP project, JGI and NCBI databases suggests that the MT-biased genes have high evolutionary rates and that the genic program for MT determination and/or signalling between mating types has a limited level of conservation. The availability of genomes of other *Pseudo-nitzschia* species and closely related taxa will allow refining this analysis.

The molecular evolution of sex (MT)-biased genes has been considered by several studies. In gonochoristic/dioecious/heterothallic systems it was observed that male-biased genes tend to evolve more quickly than female-biased genes at the protein level, suggesting that male-biased genes are under stronger selection due to male–male competition or female choice, natural selection, and/or relaxed purifying selection arising from gene dispensability or reduced functional pleiotropy (Ellegren and Parsch 2007, Ingleby *et al.* 2014, Lipinska *et al.* 2015). However, in the case of *E. siliculosus,* it was found that both male and female sex-biased genes showed accelerated rates of evolution as compared with unbiased genes explaining that the balanced rate of evolution is consistent with the low level of sexual dimorphism, which presumably provides limited scope for asymmetric sexual selection (Lipinska *et al.* 2015).

In order to identify the *P. multistriata* genes under positive selection a pair-wise comparison of the orthologs of *P. multistriata* and *P. multiseries* was performed and the Ka/Ks ratio was calculated that gives a measure of evolutionary divergence (Nekrutenko A. *et al.*, 2002 Hurst, 2002). A total of 6066 homologous pairs were found between the two species. Of these 6066 genes, 132 have a Ka/Ks value >1, indicating positive selection. In many systems which display a broad set of sex-related genes (such as multicellular organisms with macroscopic differences between the two sexes), a comparative analysis on the evolutionary rates of different sets of genes, e.g., DEG in female individuals *vs* DEG in male individuals, can provide valuable information on the evolution of mating and ecological insights. In the case of *P. multistriata* unfortunately such an analysis cannot be

considered because the validation of the DEG between the two MTs resulted in only five MT-biased genes, a number too small to perform a significant comparison on their evolutionary rates.

Among the genes having a one-to-one relationship with *P. multiseries*, two MT-biased genes were detected. *MRP1* (PSNMU-V1.4_AUG-EV-PASAV3_0024820.1) was under positive selection with a Ka/Ks of 2,13 and with a FDR of 2,83E-04 and *MRP2* with a Ka/Ks of 0,49 and with a FDR of 1,12E-06. Positive selection is common for genes involved in sexual reproduction and can contribute to maintaining reproductive isolation. Further studies will hopefully clarify whether this gene has a role in, e.g., recognizing the right mating partner, thus avoiding inter-species breeding.

Chapter 3

Pseudo-nitzschia multistriata mating type-biased genes: expression pattern during early phase of sexual reproduction, during a 24 hours L:D cycle and in sexually immature strains

3.1 Introduction

In this chapter three approaches, aimed at improving the characterization of the five MT-biased genes of *P. multistriata* identified in Chapter 2 and their role and function in MT determination, will be presented: i) expression trends of the five genes during early response of the two mating types at the beginning of the sexual phase, ii) the expression trends over a 24 hours' time course experiment to assess whether any of them was regulated by light or by the cell cycle; iii) RT-PCR analysis of the MT-biased genes in samples above the sexualisation size threshold to examine their expression profile in sexually immature strains.

In *P. multistriata,* sexualisation is induced by mixing clonal strains of opposite MT. The pairing between cells of opposite mating type is not an obvious attractive behaviour of one strain towards the other mediated by pheromones, as observed in *Seminavis robusta* (Gillard *et al.* 2013), rather cells from both mating types move actively and explore the environment until they find a cell to pair with (Scalco *et al.* 2015). For *P. multistriata,* it was shown that the onset of sex is a density-dependent event and it was suggested that a mechanism comparable to *quorum sensing* could trigger the production of sex pheromones (Scalco *et al.* 2014).

One hypothesis for the function of the five MT-biased genes is that they are linked to signaling processes mediated by chemical cues between mating types. To test this hypothesis, I analyzed the expression changes of the 91 candidate MT-biased genes (see Chapter 2) in the transcriptomic dataset gained within an experiment aimed at studying the genes that were activated in the early stages of the sexual phase. This experiment was run placing two *P. multistriata* strains each in one compartment of an apparatus that allowed free exchange of the growth medium but not physical contact between cells (Basu *et al.* under revision).

Planktonic unicellular microalgae produce a broad range of secondary metabolites that can mediate intra-cellular communication for the purpose of e.g., defence, finding a mate, switching between life cycle stages. Their ability to communicate is also strictly regulated by their capability to perceive external stimuli. Marine diatoms live in a dynamic environment exposed to periodic changes in nutrient conditions, pH, cell density, and diel light cycling. Their seasonal dominance in phytoplankton communities of marine and freshwater ecosystems suggests that they possess efficient sensory and regulatory mechanisms that allow them to respond or adapt adequately to the environmental fluctuations through the activation of specific cell pathways and cell cycle checkpoints (Ashworth *et al.* 2013, Huysman *et al.* 2013).

The S and M phases of the cell cycle are separated by two gap phases, G1 before S phase and G2 before M phase. The gap phases act as cell cycle checkpoints in response to external stimuli, such as light. Light-controlled restriction points have been identified in several diatom species, either only during the G1 phase (Chisholm *et al.* 1986, Olson *et al.* 1986, Gillard *et al.* 2008, Huysman *et al.* 2010, Huysman *et al.* 2013), or during both the G1 and G2/M phases of the cell cycle (Brzezinski *et al.* 1990). Alteration of the diel light cycle is used for synchronizing the natural cell cycle; a prolonged dark treatment causes the arrest in G1 phase and the synchronous release of the cell cycle arrest when illumination is provided again (Gillard *et al.* 2008, Huysman *et al.* 2010).

An example of regulatory system is the mechanism of pheromone production in *Seminavis robusta*, which was shown to be strictly light-dependent (Gillard *et al.* 2013). The regulation of the signalling process in *S. robusta* is a two-step system, in which cells below the SST produce cytostatic sex-inducing pheromones SIPs that reciprocally arrest the cell cycle at the G1 phase. The sex-inducing pheromone (SIP+), secreted by MT+, triggers the switch from mitosis to meiosis in MT− and induces the production of the light-dependent pheromone L-diproline that attracts male cells (Moeys *et al.* 2016). This background

information brought to analyse the MT-biased genes in 24 hours' time course experiment to understand whether their expression was regulated by light or cell cycle.

Finally, the observation reported by (Moeys *et al.* 2016), where only cells below the SST produce and perceive chemical signals, and the results of the differential expression analysis reported in Chapter 2 (section 2.3.5, Table 2.8), where the five MT- biased genes showed zero or few counts in samples above the SST, prompted to expand the analysis of the MT-related genes in sexually immature cells, i.e. above the SST.

3.2 Material and Methods

3.2.1 *Pseudo-nitzschia multistriata* 'sensing transcriptome'

In the PhD thesis of S. Patil (2014), as reported in Chapter 1 section 1.4, it was studied the early response of the two mating types at the beginning of the sexual phase in an experimental set up in which strains were physically separated but in contact through their culture medium. Samples were collected from the experimental set up and from two controls (the parental strains in mono-culture) at two time points: 2 hrs (T1=10:30 AM) and 6 hrs (T2=2:30 PM) after the co-culture was started (Fig. 3.1). A transcriptome was produced for the analysis of differentially expressed genes; this transcriptome is defined as 'sensing transcriptome' from here onwards because it provided information on the signalling and metabolic responses of the two mating types that perceived each other.

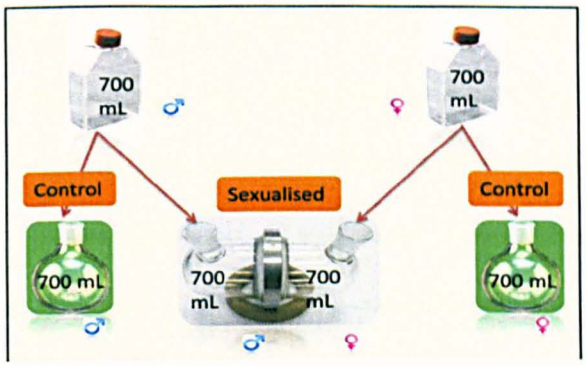


Figure 3.1: The apparatus used to generate the 'sensing transcriptome' (Patil, 2014). The double glass flasks are separated by a membrane filter of hydrophilic polyvinylidene fluoride (PVDF) with 0.22 μm pore size.

The experiment was conducted on two biological replicas, i.e. two pair of strains with different mating type, for a total of 16 samples (Table 3.1). Strains B938 (MT+) and B857 (MT-) were used in experiment 1, strains B856 (MT+) and B939 (MT-) were used in experiment 2.

Table 3 1: List of the 16 samples used to generate the sensing transcriptome; A and B mark the two different experiments (Patil 2014).

Sample name	Mating types	Time point
B938 Control A	+	T1 10:30 AM
B938 Sexualised A	+	T1 10:30 AM
B857 Control A	-	T1 10:30 AM
B857 Sexualised A	-	T1 10:30 AM
B856 Control B	+	T1 10:30AM
B856 Sexualised B	+	T1 10:30 AM
B939 Control B	-	T1 10:30AM
B939 Sexualised B	-	T1 10:30AM
B938 Control A	+	T2 2:30 PM
B938 Sexualised A	+	T2 2:30 PM
B857 Control A	-	T2 2:30 PM
B857 Sexualised A	-	T2 2:30 PM
B856 Control B	+	T2 2:30 PM
B856 Sexualised B	+	T2 2:30 PM
B939 Control B	-	T2 2:30 PM
B939 Sexualised B	-	T2 2:30 PM

The expression values, as normalized counts (CPM=counts per million), obtained from the analysis of the 'sensing transcriptome' has been used to visualize the expression trend of the 91 putative MT-biased genes (see Chapter 2, section 2.3.2) within this dataset.

3.2.2 Set up of the synchronization protocol

To verify whether *Pseudo-nitzschia multistriata* cells could be dark-synchronized, strains were incubated in the dark for 36 hours. The experiment was conducted on strain B856 (MT+). Two of three subcultures were tested for synchronization, while the other one was

used as control. The three replicate cultures were grown in culture flasks filled with 25 ml of F/2 medium, with a semi-continuous protocol (MacIntyre and Cullen 2005) under standard growth conditions (temperature of 18 °C, irradiance of 100 μmol photons m^{-1} sec^{-1}, and 12L:12D h photoperiod). Exponentially growing cultures were diluted to achieve a starting cell concentration of 70,000 cell·ml^{-1} and this cycle was repeated until a constant growth rate was achieved. At this point, the replicate samples were diluted and two of them were kept in the dark for 36 hours. After dark incubation, cultures were exposed to standard growth conditions and sampled every two hours for the following 12 h for a total of seven time points. The control was kept at the standard growth conditions and was sampled at the same time points of the dark-synchronized strains. For each replica, 5 ml of culture were collected at each time point, placed in an Eppendorf vial and fixed with formaldehyde (1.6% final concentration).

In order to visualize the nuclei, all samples were analysed as follows: cells were stained with 1 μl·ml^{-1} DAPI working solution (0, 5 mg·ml^{-1}) for 15'. The stained samples were placed in an Utermöhl sedimentation chamber and examined with a Zeiss Axiovert 200 epifluorescence microscope equipped with the filter FS09 (excitation, 450 to 490 nm; emission, 515 nm) at 400x magnification. The dividing cells (with two nuclei) and non-dividing cells (with one nucleus) (Fig. 3.2) were enumerated in ten random fields. Cell counts were converted to cell·ml^{-1}.

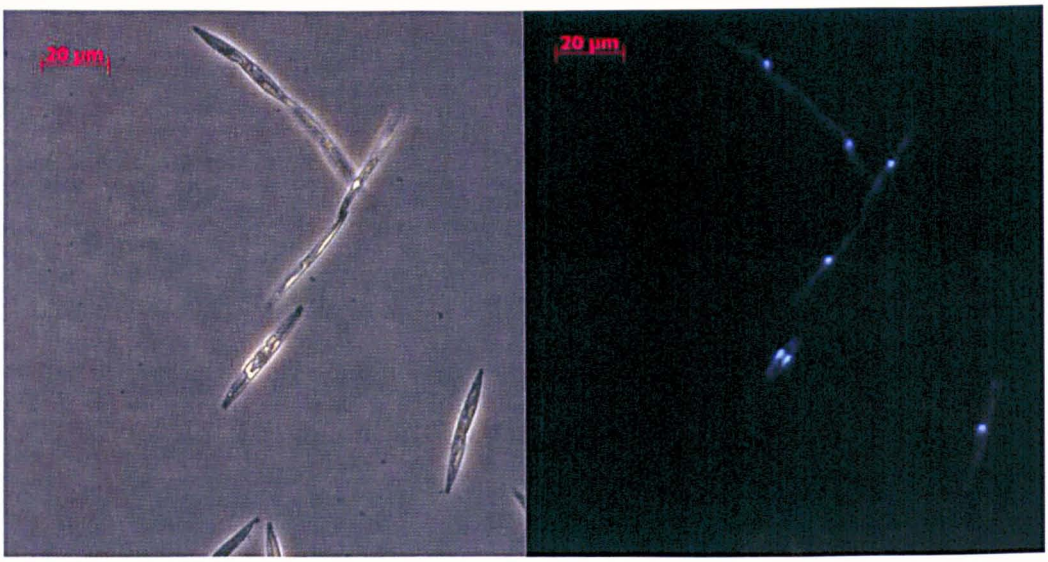


Figure 3.2: Photographs of synchronized cells of *P. multistriata* stained with DAPI (Zeiss Axiovert 200 epifluorescence microscope). Left panel, bright field image. Right panel, fluorescence image, blue for DAPI staining.

3.2.3 Cultures for the 24 h time course experiment

The strains of *Pseudo-nitzschia multistriata* used for the experiment were isolated at the LTER-MC station in Gulf of Naples in 2013 or obtained by crosses carried out in the lab (Table 3.2).

Table 3.2: Strains of *P. multistriata* used for the 24 h time course experiment. For each strain are reported: the strain code, the mating type, the average apical length and the origin of the strains.

Strain code	**Mating type (MT)**	**Apical length**	**Origin**
MVR171.1	MT-	15 µm	Wilde type
SH20	MT-	15 µm	F1 (Sy776*B935)
LV133	MT-	34 µm	F1 (B854*MVR1041.1)
LV96	MT+	38 µm	F1 (B854*MVR1041.1)
LV130	MT+	28 µm	F1 (B854*MVR1041.1)
LV131	MT+	30 µm	F1 (B854*MVR1041.1)

The cultures were grown in F/2 culture medium (Guillard 1975) prepared with oligotrophic seawater, as illustrated in Chapter 2 (section 2.2.7). Strains were maintained in a growth chamber at a temperature of 18 °C, a photoperiod of 12:12 h Light:Dark, and a photon flux density of 50-60 µmol photons $m^{-2}s^{-1}$ provided by cool white fluorescent tubes (TLD 36W/950, Philips, Amsterdam, Nederland). The selected strains represented biological triplicates for each mating type.

All strains were tested for cross efficiency and they resulted capable of producing a good percentage of sexual stages (at least 20%) when crossed with strains of opposite mating type. For mating experiment see Chapter 2 (section 2.2.8).

3.2.4 Experimental design and culturing conditions

To investigate the natural expression trend of the MT-biased genes during a 24 h cell cycle (12L:12D h) a time course experiment was conducted. I selected only the four MT-biased genes that turned out to be changing according to time in the analysis of the 'sensing transcriptome'. The three pairs of MT+ and MT- strains were grown exponentially (180-200·10^{-3} cells·mL^{-1}) at 80-100 µmol photons·m^{-2}·s^{-1} light, 12:12 L:D photoperiod and 20 ^{0}C temperature in 2L F/2 +Si medium. The cell cycle of exponentially growing cultures was synchronized by incubating the cultures in the dark for 36 hours. The synchronized cultures, still in dark, were diluted to 80-100·10^{-3} cells·ml^{-1} concentration. The sampling for the estimation of cell concentration and the subsequent dilution procedure were performed at dim red light illumination. The 2 L cultures were subsequently split in nine 200 ml subsamples; also this procedure was carried out at dim red light condition. Cultures were then brought back to light (80-100 µmol photons m^{-2} s^{-1} light) and, in the next 24 h, at each sampling point one aliquot was taken.

In order to verify that the mating efficiency was not affected by dark incubation, three cross tests were prepared in 6 well culture plates (Costar tissue culture plates, Corning Inc.,

NY, USA) three hours before the end of the dark phase. One plate was kept at the same growth condition of the experimental plan (i.e. they experienced the same L:D cycle as the experimental strains), while the other two plates were kept in complete darkness for other 12 and 72 hours of dark, respectively.

3.2.5 Sampling

Nine time points were sampled, for a total of 54 samples; at each time point, one bottle of 200 ml for each MT- and MT+ strains was collected. Sampling was performed every two or three hours along the 24 h L:D cycle, five time points during the light phase and four time points during the dark phase (Fig. 3.3).

At each time point, the 200 ml of each culture were collected as following:

- 150 ml for filtration and RNA extraction;
- 40 ml was centrifuged at 2906 g (Centrifuge 5810 R, Eppendorf), 4 ^{0}C, for 20 min, re-suspended in 1 ml ice-cold methanol and stored at -20 ^{0}C for DNA content analysis using flow cytometry;
- 10 ml was preserved with formaldehyde (1.6 % final concentration) at 4 °C for future observations in light microscopy.

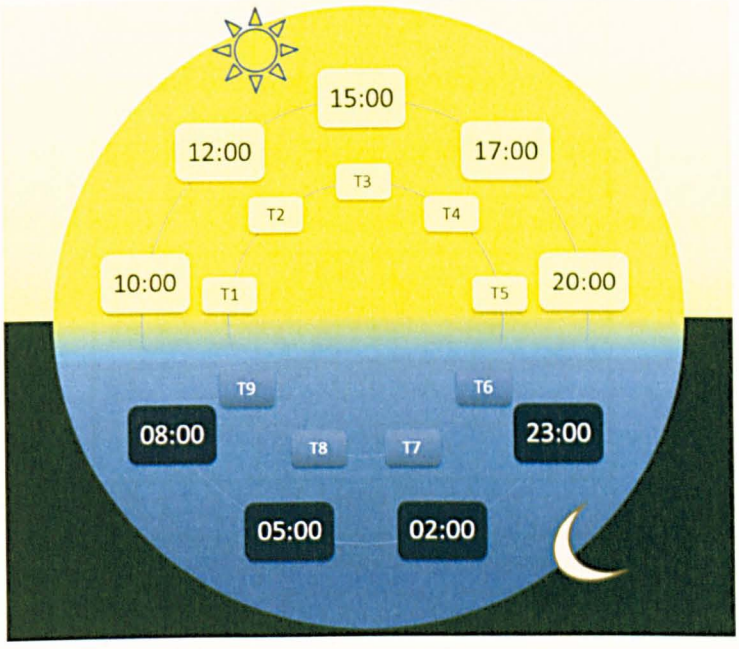


Figure 3.3: Time points of the 24 hours' time course experiment. T1-T5 were collected during the light phase while T6-T9 during the dark phase.

3.2.6 Samples filtration, RNA extraction and cDNA preparation

For samples filtration, RNA extraction and cDNA preparation see Chapter 2, section 2.2.9. The only difference in the protocol of the present experiment was in the cDNA preparation, where 500 ng (and not 1 µg as in Chapter 2) of total RNA extracted was used with the QuantiTect® Reverse Transcription Kit (Qiagen).

3.2.7 Flow cytometry

To observe cell cycle progression throughout the 24 h time period and to verify the success of the synchronization protocol, samples were analyzed for DNA content in flow cytometry. From the -20 °C frozen samples, methanol was removed by centrifugation at 2655 g (Centrifuge 5417 R, Eppendorf), for 5 min, at 4 °C. Pellets were washed with and re-suspended in 1mL Tris-EDTA buffer (pH 8). DNase free RNase (300 µg mL-1) was added to the samples and they were incubated for 45 min at room temperature. Then

1:10,000 dilution of SYBR Green stock (SYBR® Green I Nucleic Acid Gel Stain - 10,000X concentrates in DMSO, Invitrogen™) was added to 1 ml of the cell suspension, which was then mixed briefly on a vortex. Samples were incubated in the dark for 10 to 15 min to allow proper staining and then analysed with BD FACS Calibure flow cytometer for DNA content. The flow cytometry analyses were performed in collaboration with Dr. Raffaella Casotti, at Stazione Zoologica Anton Dohrn.

3.2.8 qRT-PCR validations

Quantitative real time PCR analyses were performed on six time points of the experiment on all the samples (Tab. 3.3) for a total of 36 samples. The aim was to quantify the expression levels of the four MT-related genes resulted from the differential expression analysis illustrated in Chapter 2.

Table 3.3: List of the samples on which PCR validations were conducted. The following information is reported: time point, sample code (composed, respectively, by time point, strain code and mating type).

Time point	**Sample code**
T1	T1 MVR171.1 Pm-
T1	T1 SH20 Pm-
T1	T1 LV133 Pm-
T1	T1 LV96 Pm+
T1	T1 LV130 Pm+
T1	T1 LV131 Pm+
T2	T2 MVR171.1 Pm-
T2	T2 SH20 Pm-
T2	T2 LV133 Pm-
T2	T2 LV96 Pm+
T2	T2 LV130 Pm+
T2	T2 LV131 Pm+

T4	T4 MVR171.1 Pm-
T4	T4 SH20 Pm-
T4	T4 LV133 Pm-
T4	T4 LV96 Pm+
T4	T4 LV130 Pm+
T4	T4 LV131 Pm+
T5	T4 MVR171.1 Pm-
T5	T5 SH20 Pm-
T5	T5 LV133 Pm-
T5	T5 LV96 Pm+
T5	T5 LV130 Pm+
T5	T5 LV131 Pm+
T7	T7 MVR171.1 Pm-
T7	T7 SH20 Pm-
T7	T7 LV133 Pm-
T7	T7 LV96 Pm+
T7	T7 LV130 Pm+
T7	T7 LV131 Pm+
T9	T9 MVR171.1 Pm-
T9	T9 SH20 Pm-
T9	T9 LV133 Pm-
T9	T9 LV96 Pm+
T9	T9 LV130 Pm+
T9	T9 LV131 Pm+

Information on the four primer pairs of the target genes are illustrated in Chapter 2, section 2.2.10.

3.2.9 qRT-PCR data analysis and statistics

For *REST – qRT-PCR data analysis* methods please refer to Chapter 2, section 2.2.10. To validate the expression rates of the four targets within this experimental setup it was necessary to consider a few relevant constrains: i) there was no reference condition with which to compare the different time points; ii) there was more than one variable to consider (time, mating type and strain-specific variability) that made it difficult to analyse the data using REST (Pfaffl et al., 2002).

There are several methods to present relative gene expression so, to overcome the constrains mentioned above, I decided to combine REST analysis with an additional approach:

Comparative quantification with ΔCT method, where ΔCT, using raw (non-normalized) gene expression values, was equal to the difference in threshold cycles (CT) for target and reference genes (CT,t - CT,r). The CT,r value (CT of the reference gene) was obtained by calculating an arithmetic mean for the three reference genes (*CDK,TBP,COPA*) (Schmittgen and Livak 2008). The result was transformed in $2^{-\Delta CT}$ ($\log_{10}$ normalized) and coupled with factorial ANOVA validations to determine whether the differences between MTs and time were statistically significant.

STATISTICA 7 software was used to perform statistical analyses (Hilbe 2007). Levene's Test for Homogeneity of Variances was performed to check the homogeneity of the data so to decide whether to consider significant the ANOVA p-value <0.05. Factorial ANOVA was then applied to the all data set of $2^{-\Delta CT}$ values for each gene and then examined with a Student–Newman–Keuls test (SNK). This test is a stepwise multiple comparisons procedure used to identify sample average values that are significantly different from each other, member of the *post hoc* analyses that usually concern with finding patterns and/or relationships between subgroups of sampled populations that would otherwise remain undetected.

3.2.10 MT-biased genes in strains above the sexualisation size threshold

The four MT-biased genes were analysed on the cDNA of six strains above the sexualization size threshold (>SST) of *P. multistriata* by RT-PCR.

Among the six strains, three were proved to be MT+ after size reduction and three were proved to be MT-. The determination of mating type was carried out by crossing the strains, when they have reached the size < SST, with reference strains of known mating type (see Chapter 2, section 2.2.8).

Table 3.4: List of primers of the MT-biased genes validated through RT-PCR. Reported the gene code, primer name, primer sequences and amplicon size.

Gene Code	Primers pairs	Sequence	Amplicon size (bp)
MRP1	0.00+F 0.00+R	5'-GTATGGCGCTCACCACTTC-3' 5'-CGTCTTCGACTGCGTCTTC-3'	156
MRP2	127.15 F3 127.15 R3	5'-CCTCCGAATATGGATACATG-3' 5'-GAGCTAAACATCGTGACACC-3'	194
MRM1	47507F 47507R	5'-CCCCTACAAGCTCTTTGATTTG-3' 5'-GAAATTGTGGTGCCCAAAG-3'	160
MRM2	46228F 46228R	5'-CCACCGAACTAGGCAACTGTC-3' 5'-GGCACAGAACCCGTCAAC-3'	139

Table 3.5: Strains of *P. multistriata* above the SST used for the validation. For each strain are reported: the strain code, the mating type, the average apical length and the RNA extraction date.

Strain code	**MT**	**Average apical length**
LV93 B	MT+	91.5µm
LV128 B	MT+	85.4µm
LV149 B	MT+	91.5µm
LV92 B	MT-	91.5µm
LV98 B	MT-	85.4µm
LV129 B	MT-	91.5µm

The quality check of the cDNA was conducted by amplifying the constitutive H4 gene with Fw/Rv primer pairs (Fig. 3.4) and on the constitutive TUB A by amplifying a fragment of 1Kb containing one intron with primers TUB A Fw intron: 5'-

CGAGAGTAACCTTTAAATGCCAAG-3' and TUB A Pm rv: TUB A Pm rv 5'-GACGACATCTCCACGGTAC-3' (Fig. 3.5).

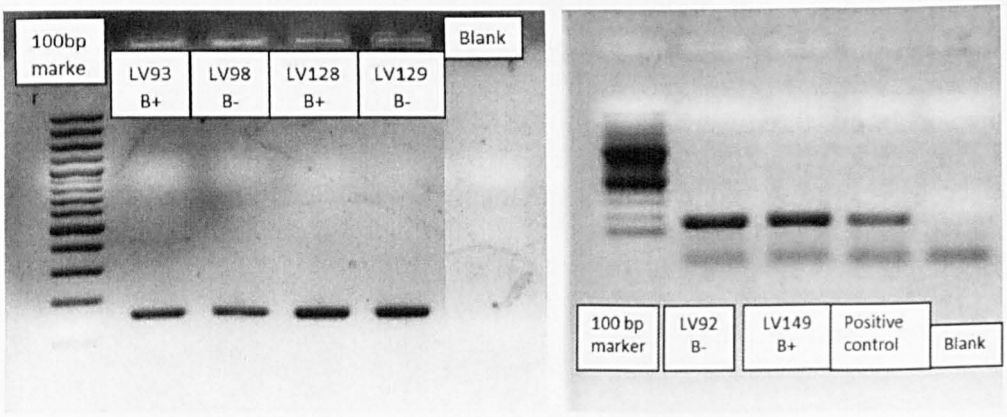


Figure 3.4: RT-PCR, quality check of the six cDNA samples with the constitutive H4 gene.

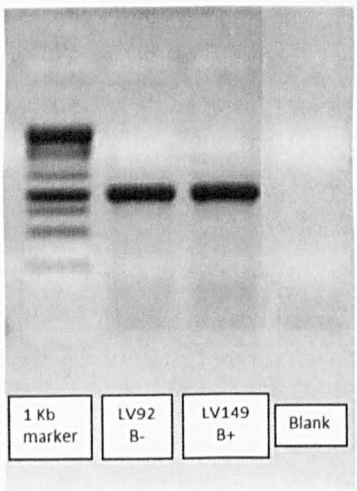


Figure 3.5: RT-PCR, quality check of two random cDNA samples with the constitutive TUB A gene.

As positive controls of the RT-PCR two strains <SST were selected, known to express the target genes. One MT+ (B937 sexualised T2) for the MT+ biased genes (*MRP1, MRP2, MRP3*), and one MT- (B936 sexualised T2) as positive control of the MT- biased genes (*MRM1, MRM2*).

3.3 Results

3.3.1 Expression patterns of the MT-biased genes in the early phase of sexual reproduction

To understand weather some of the candidate MT-biased genes of *P. multistriata* were involved in the early phase of mating recognition during sexual reproduction, I looked at their expression trend in the *P. multistriata* 'sensing transcriptome'. The differential expression analysis conducted on MT+ and MT- strains of *P. multistriata* transcriptome (see Chapter 2, section 2.3.2) resulted in a list of 91 putative MT-biased transcripts. I checked their expression profile in the available dataset of the 'sensing transcriptome'. Only 72 genes were recorded in the dataset. The Heat Map reported in Figure 3.6 represents the expression levels of the selected transcripts for all the 16 samples.

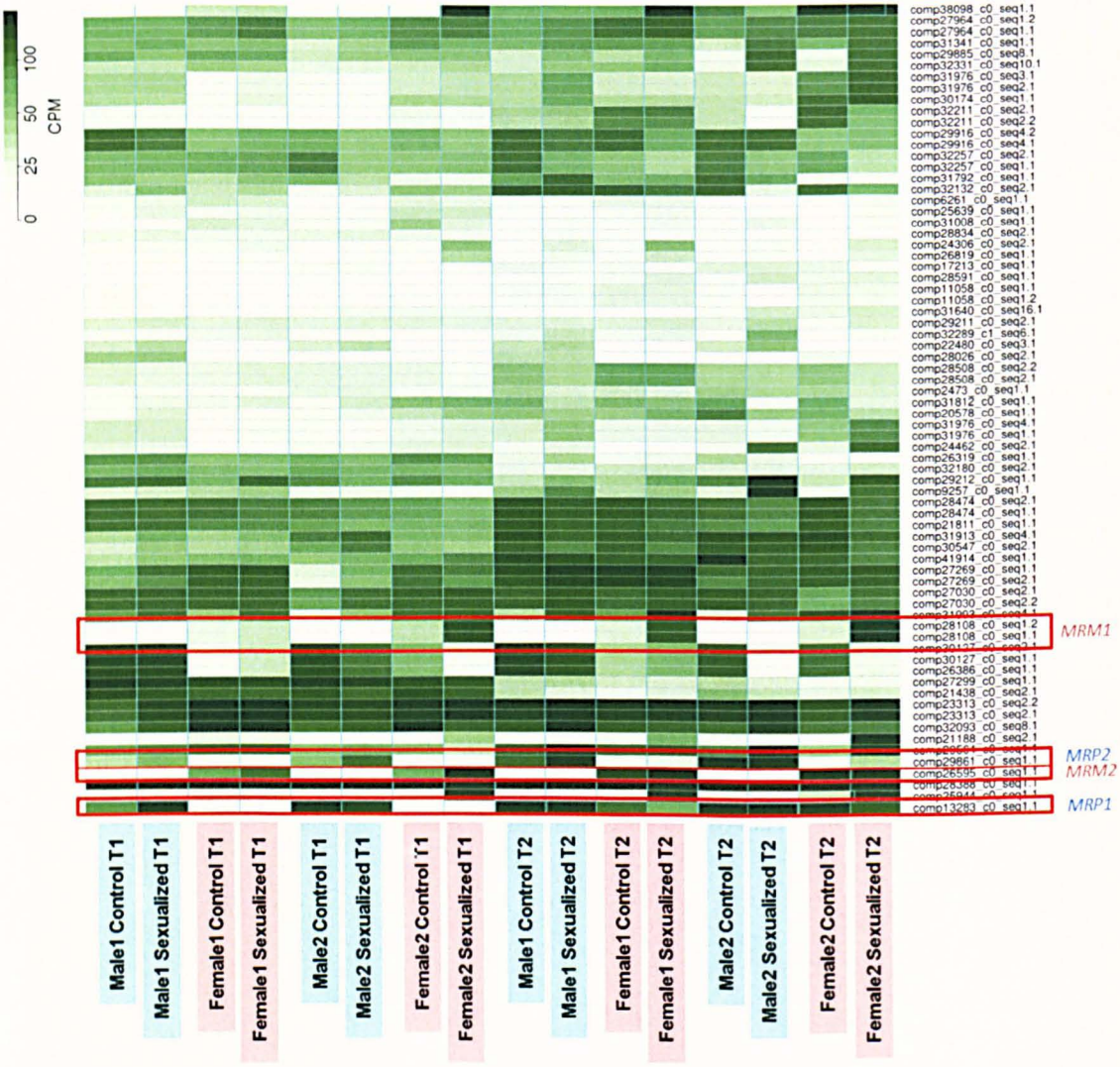


Figure 3.6: Expression profile of the putative MT-biased transcripts within the 'sensing transcriptome'in CPM (counts per million). The four mating type related transcripts are marked by a red frame. *MRM1* shows two isoforms.

The analysis of the normalized counts (CPM) revealed that four out of five MT-biased genes showed an increase of expression in the sexualised samples against the controls (Fig. 3.7). Only *MRP3* was not changing its expression trend in relation to the sexualised phase of the samples.

Gene code	Transcript ID	B938_CL. early	B938_SL. early	B856_CL. early	B856_SL. early	B857_CL. early	B857_SL. early	B939_CL. early	B939_SL. early
MRP1	comp13283	42	168	204	117	0	0	0	0
MRP2	comp29861	18	31	27	58	4	3	1	2
MRP3	comp20279	7	9	14	11	0	0	0	0
MRM1	comp28108	0	0	0	0	10	16	18	94
MRM2	comp26595	0	0	0	1	37	52	34	307
Gene code	**Transcript ID**	**B938_CL. late**	**B938_SL. late**	**B856_CL. late**	**B856_SL. late**	**B857_CL. late**	**B857_SL. late**	**B939_CL. late**	**B939_SL. late**
MRP1	comp13283	749	2417	1589	78084	52	32	319	85
MRP2	comp29861	59	142	212	963	5	8	6	21
MRP3	comp20279	7	6	11	21	0	0	0	0
MRM1	comp28108	0	0	0	0	12	80	14	210
MRM2	comp26595	1	1	1	1	104	557	251	1551

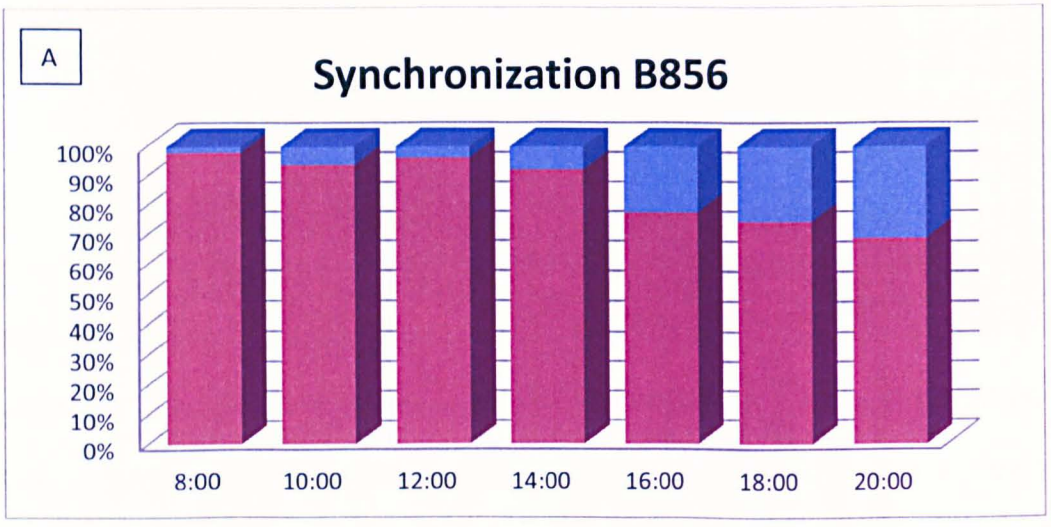
Figure 3.7: Normalized counts of the five MT related genes within the sensing transcriptome. Reported are: the gene code, the transcript ID and the sample code (e.g., B938: strain code, CL./SL.: control/sexualised phase, early/late: T1/T2). The sexualised samples are highlighted in dark blue or dark pink.

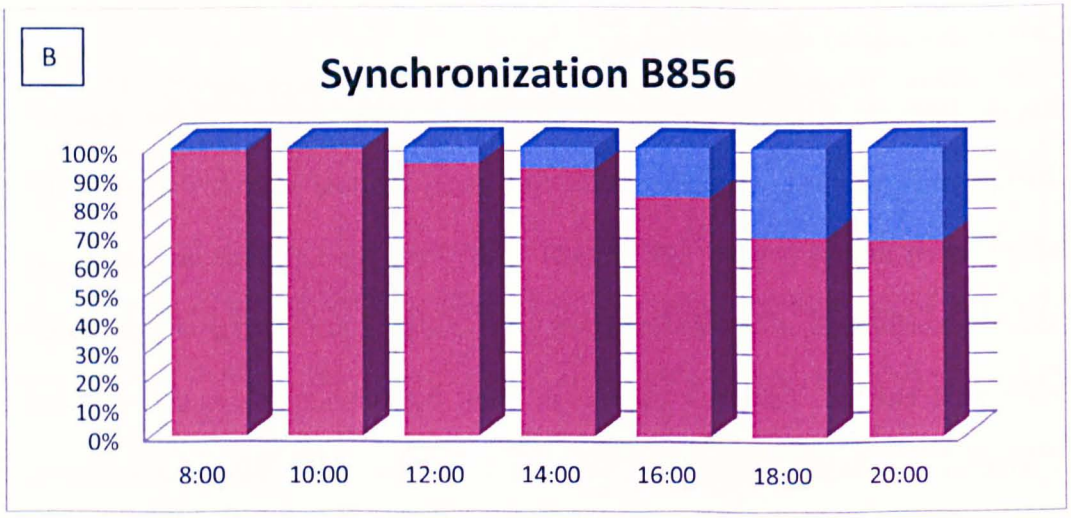
The expression levels of *MRP1*, *MRP2*, *MRM1* and *MRM2*, but not that of *MRP3*, were increasing in the samples collected at the second time point (late), in respect to the first time point (early) (Fig. 3.7). All the four transcripts increased in the sexualised samples from the early to the late time point (T1 < T2). Within controls, *MRP1* and *MRM2* showed a significant increase in expression rates between the two time points, while *MRP2* and *MRM1* were only slightly changing.

Summing up, from the differential expression analysis conducted on the transcriptome of *P. mulistriata* MT+ and MT- vegetative cells, five genes resulted to be MT-biased. Three were MT+ biased and two were MT- biased. Analysing their behaviour during early beginning of sexualisation, it was detected that *MRP1*, *MRP2*, *MRM1* and *MRM2* not only were MT-biased but also that their expression trend was higher in sexualised samples against controls and that in sexualised samples their expression increased in a time-dependent manner.

3.3.2 Set up of the synchronization protocol

The dark synchronization of *Pseudo-nitzschia multistriata* resulted in very low percentages of dividing cells in the two replicates synchronized samples, above all for the first 6 h after re-illumination, evidencing the cell arrest in G1 phase. In Figure 3.8 are plotted the percentages of dividing and non-dividing cells.





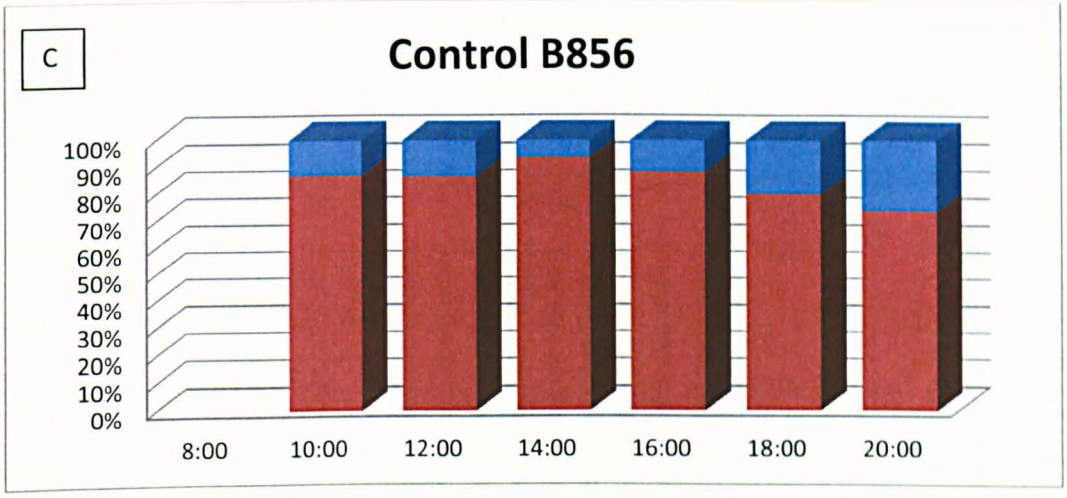


Figure 3.8: The percentage of dividing (blue) and non-dividing (red) cells in the seven sampling points after dark synchronization; panels A and B: two biological replicates, panel C: the non-synchronized control. The first sample of the control was lost.

After 36 hrs of dark incubation (at the time at which the dark treatment was released), about 82% of cells arrested their cell cycle in G1 phase whereas in control condition, 68% cells were already in phase G2 (0 h, 8:00 AM). With progressing time cells continued to move into G2+M phase and at 9 h (5:00 PM) after light re-illumination, about 59 and 52% of dark synchronized and control cells, respectively, were into G2 phase. An interesting point was the moderate degree of natural synchronization exhibited by the control culture.

3.3.3 Cross efficiency and flow cytometric analysis of the 24 h time course experiment

In parallel to the experiment, a series of crosses were carried out to test if the mating efficiency was affected by prolonged dark incubation. The strains crossed after 36 h of dark incubation were perfectly able to mate once light was provided, as also the ones kept in dark for additional 12 h. On the contrary, the cross made after 72 h of dark was not able to produce sexual stages. Strains SH20 and LV130 were used as representative for the flow cytometric analysis of DNA content. The results confirmed the effective synchronization

of the experimental cultures and the cell cycle arrest in G1. The cell cycle progression is illustrated by the percentage of cells in G1 *vs* S+G2+M in Figure 3.9.

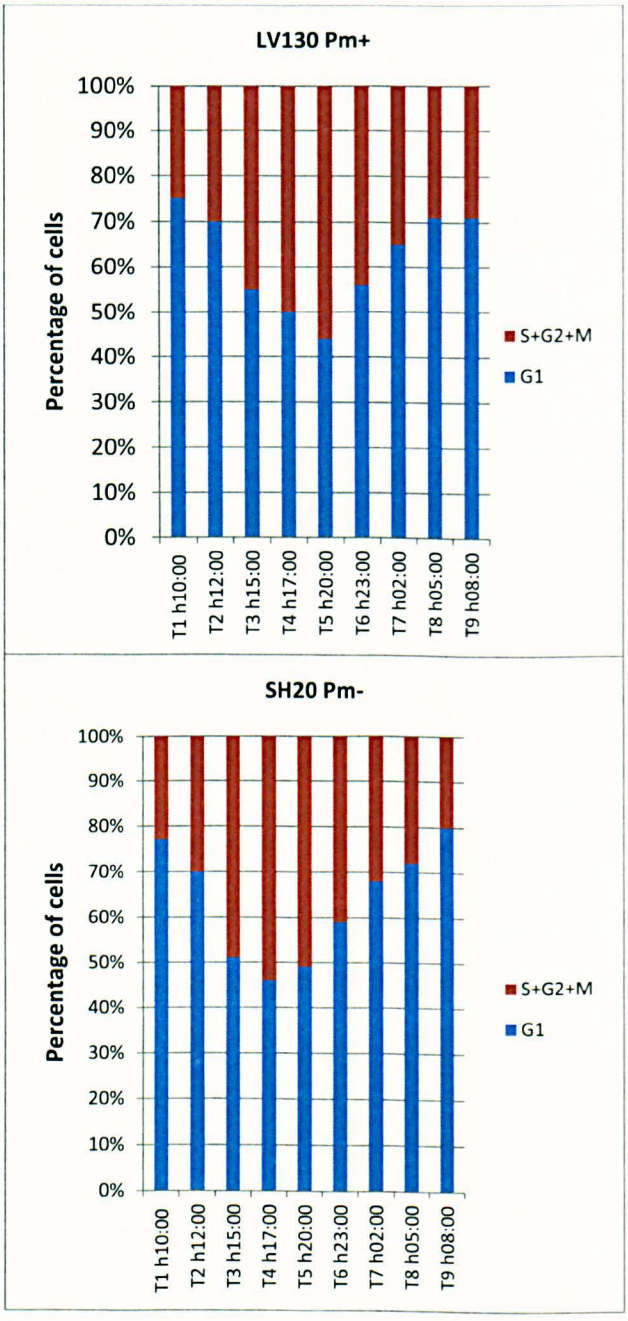


Figure 3.9: Flow cytometric analysis of DNA content. In the upper panel: DNA content of LV130 (MT+) during 24 h cycle. In the lower panel: DNA content of SH20 (MT-) during 24 h cycle. In blue the % of cells in G1, in red the % of cells in S+G2+M.

3.3.4 CT study and REST analysis for the reference and target genes used in the 24 h time course experiment

The 24 h time course experiment, designed to detect gene expression variation of the four MT-biased genes, was performed in triplicate for each mating type (3MT- and 3MT+). Six time points were considered for the qRT-PCR validations (Table 3.3). The T5 sample of strain SH20 Pm- was not considered in the analysis because its RNA was of low quality. The total number of samples analysed was: 6 samples x 6 time points -1 missing sample = 35 samples, i.e. 35 CT values for each gene analysed. To have a clear picture of the data set obtained with the 24 h time course experiment and to decide the right strategy for analysing it, a CT study of all the genes tested in qRT-PCR was performed. The CT is defined as the PCR cycle at which the fluorescent signal of the reporter dye crosses an arbitrarily placed threshold. By presenting data as the CT, one ensures that the PCR is in the exponential phase of amplification. The numerical value of the CT is inversely related to the amount of amplicon in the reaction, i.e., the lower the CT, the larger the amount of amplicon.

As expected, the reference genes showed a small range of CT distribution within all the data set, also if tested separately for the MT+ and the MT- samples.

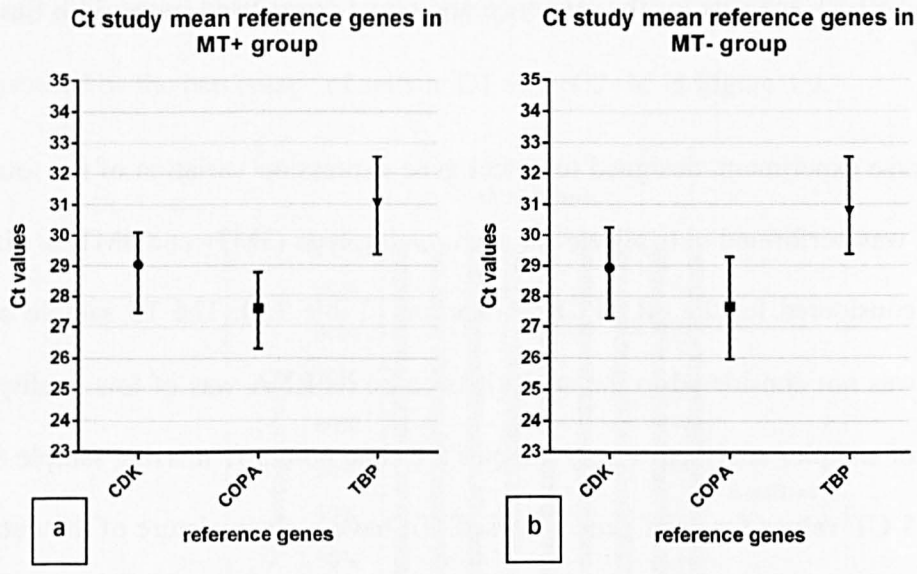


Figure 3.10: Expression levels of the reference genes in samples of different mating type. (a): CT values in MT+ samples (LV96 Pm+, LV130 Pm+, LV131 Pm+), (b) CT values in MT- samples (MVR171.1 Pm-, SH20 Pm-, LV133 Pm-), taking into account six time points during the 24 h cycle. Values are expressed as qRT-PCR cycle threshold (CT values). The lines represent the range of the average CT values measured for the 6 time points; the average CT values are represented with a symbol.

From Fig. 3.10 we can see that gene expression variations are:

in the MT+ samples group (Fig. 3.10a):

- 2.65 ΔCT CDK
- 2.51 ΔCT COPA
- 3.18 ΔCT TBP

in the MT- samples group (Fig. 3.10b):

- 2.98 ΔCT CDK
- 3.33 ΔCT COPA
- 3.14 ΔCT TBP

The COPA ΔCT of 3.33 in the MT- samples (Fig. 3.8b) was suggesting an higher variation of this reference gene. This assumption was validated by REST analysis (Fig. 3.11).

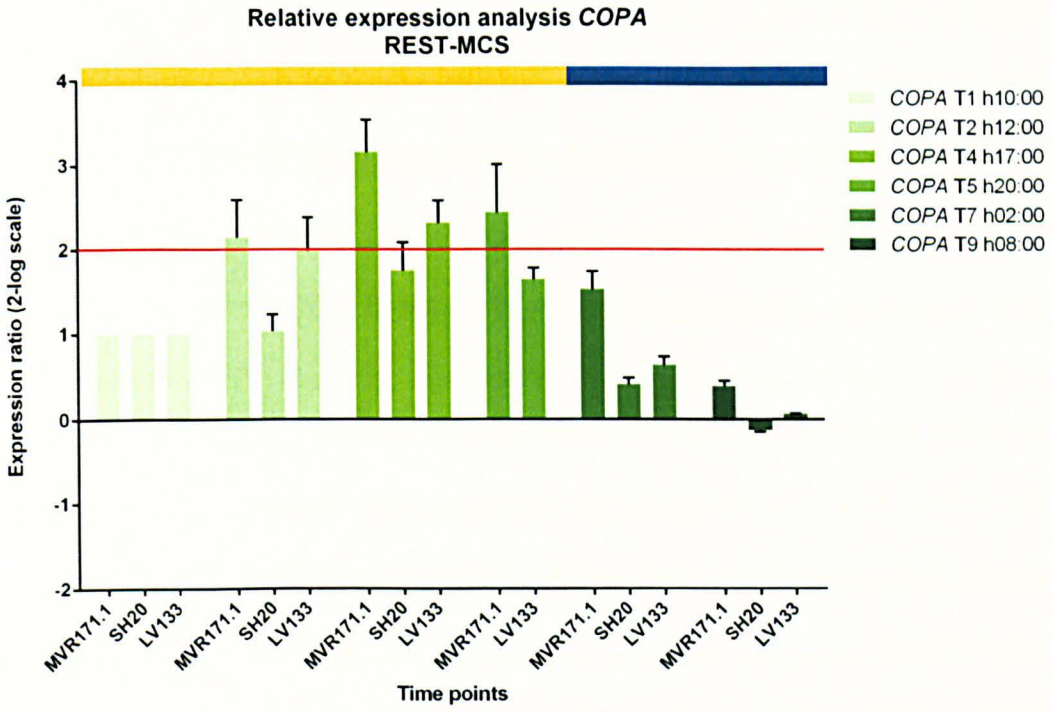


Figure 3.11: REST analysis of *COPA* obtained by fixing T1 as reference condition. Values are normalized against two reference genes CDK A and TBP.

COPA, in this experimental condition, was showing a significant up-regulation in MT- samples during T4 and T5; so it cannot be considered anymore a reference gene for MT- in this experimental dataset. It can be concluded that the most stable genes to consider for normalization for the MT- group are CDK and TBP while, all three genes (CDK, COPA and TBP) could be considered for the MT+ group. To draw such conclusions, I referred also to Adelfi *et al.*, (2013) and Siaut *et al.*, (2007) since these papers provide evidence that all genes are regulated under some conditions and, probably, there is no universal reference gene with a constant expression in all organisms (Kubista *et al.* 2006).

The CT study showed that MT-bias is conserved for all the four genes along the 24 h cycle. The reliability of the expression variation observed among the six time points analysed has been validated after normalization on the reference genes by REST. The expression variation of only *MRP1* and *MRP2* resulted significant. Both presented low expression rates at the beginning of the experiment (T1) than tended to exponentially increase till T5

and remained constant until T9. On the contrary, *MRM1* and *MRM2* did not show any expression variation along the 24 h course. Detailed graphs of the CT study and REST analysis, for both reference and target genes, are presented in APPENDIX D.

3.3.5 ΔCT comparative quantification method and statistical analysis

The expression profile of the four MT-biased genes is presented as -ΔCT in the heatmap reported in Fig. 3.12. The heatmap was generated with Pheatmap (R package) that automatically apply log transformation, and the script was modified to remove the default setting of clustering.

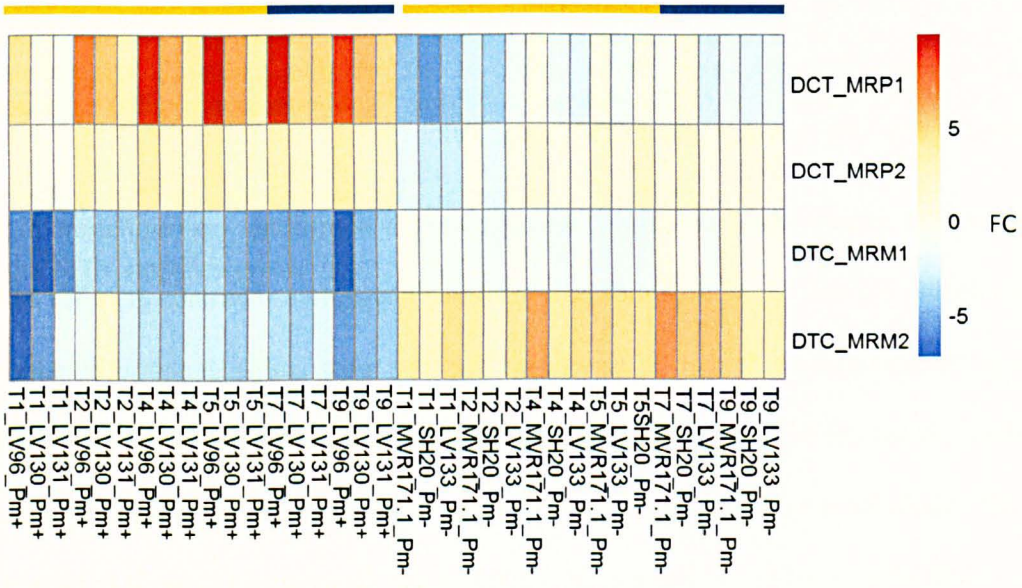


Figure 3.12: Heatmap of the expression profile of the four MT-biased genes. Fold change (FC) data $\log_{10}$ transformed.

ΔCT for all the 35 samples, calculated for the four MT-biased genes, was transformed in $2^{-\Delta CT}$ ($\log_{10}$ normalized). The ANOVA was then performed to test their significance. The factorial ANOVA was performed taking into account two factors: mating type and time, and its "Univariate Tests of Significance" tested three effects: mating type, time and mating type*time.

The p-value resulted to be significant ($p<0.05$ or $p<0.01$ if Leven's test was showing NON homogeneity of the variance) for:

- Mating type in all the four MT related genes; meaning that the differences between mating types all along the 24 hours' time course was always significant. It was confirmed also by SNK test, except for *MRP2* in T4 and T5.
- Time in *MRP1*; the only significantly differing time point was T1 in the MT+ subgroups of *MRP1*.

The effect mating type x time instead was never significant.

Comparing these results with those obtained from REST analysis, I can conclude that the only significant variation within a 24 h time course for the four genes related to MT was visible in *MRP1* and only at T1. So, three out of four candidates are not related to light or to any of the cell cycle phases. What remains to be understood is the low expression of *MRP1* at 10:00 a.m.

3.3.6 MT-biased genes in strains above the sexualisation size threshold

It is known that *P. multistriata* strains above the sexualisation size threshold (>SST) are unable to undergo sexual reproduction. Indeed, the five MT-biased genes were weakly or not expressed at all in strains above the SST, as evidenced from the results of the transcriptomics analysis (Table 3.6). To verify these observations, RT-PCR validations were performed on six samples >SST.

Table 3.6: MT-biased transcript ID, the assigned gene name and the normalized counts provided for S1+ = Sy373 small, S2+ = B856 small, L2+ = B856 large, S1- = Sy379 small, S2- = B857 small, L2- = B857 large (see Table 2.9).

Transcript ID	**Gene name**	**S1-**	**S2-**	**L2-**	**S1+**	**S2+**	**L2+**
comp13283_c0_seq1	***MRP1***	0.70	0.70	0.17	1068.26	1621.92	0.75

comp29861_c0_seq1	***MRP2***	2.66	13.76	9.31	144.86	151.75	15.74
comp20279_c0_seq4	***MRP3***	0.00	0.00	0.00	9.89	7.14	0.00
comp28108_c0_seq1	***MRM1***	3.50	26.67	0.05	0.00	0.02	0.00
comp26595_c0_seq1	***MRM2***	40.76	202.51	0.14	0.13	0.17	0.00

[(*) *MRP3* has not yet been tested because it was later discovered to be an MT-biased gene.]

The RT-PCR validation supported the working hypothesis for *MRP1*, which was not expressed in samples of both MT > SST (Fig. 3.13). The validation confirmed the prediction of the RNA-seq also for *MRP2* (Fig. 3.14). The gene was expressed also in strains >SST but with lower expression as compared to the positive control and without any difference among MTs.

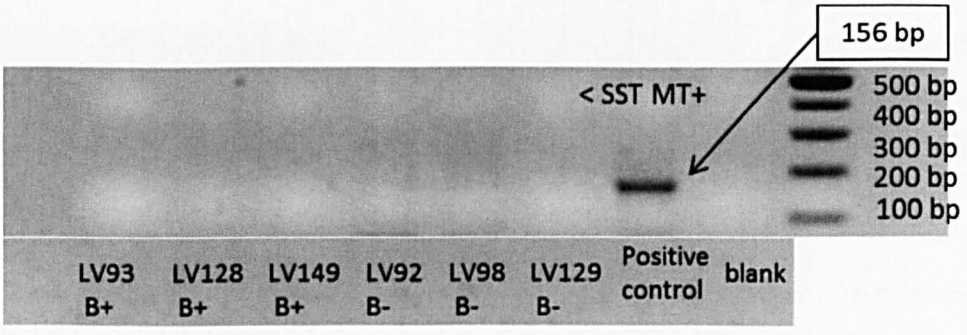


Figure 3.13: RT-PCR of *MRP1* gene on strains >SST. The positive control is a MT+ <SST sample where it is known that the gene was expressed.

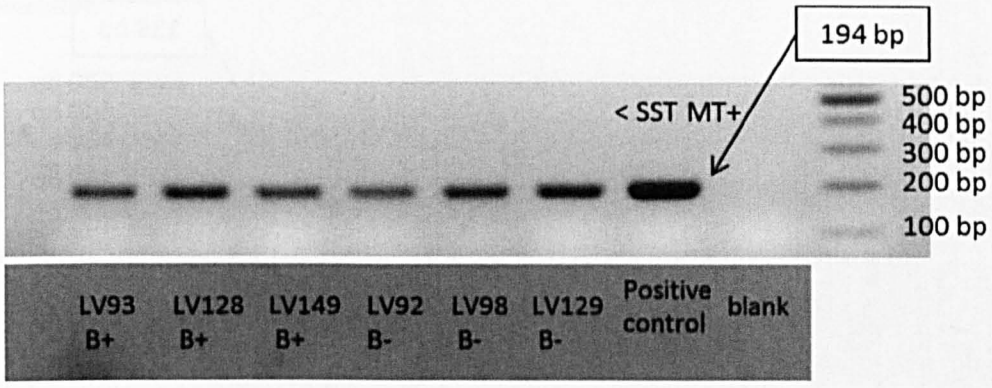


Figure 3.14: RT-PCR of *MRP2* gene on strains >SST. The positive control is a <SST MT+ sample where it is known that the gene was expressed.

Predictions of the expression pattern for *MRM1* and *MRM2* were partially supported by the RT-PCR analysis. *MRM1* presented a very faint band in samples LV92 B- and LV129 B-; however the comparison with the positive control showed the marked difference between the expression levels between the two size categories (Fig. 3.15). Also *MRM2* presents a very faint band in samples LV92 B- and LV98 B-, but also in this case the comparison with the positive control shows an evident difference in expression between >SST and <SST, suggesting that this gene is primarily expressed in cells below SST.

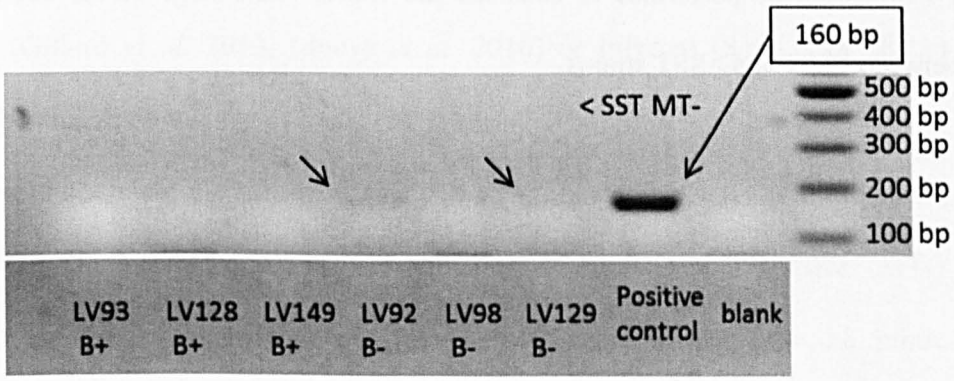


Figure 3.15:RT-PCR for the *MRM1* gene on strains >SST. The positive control is a <SST MT- sample where it is known that the gene was expressed. The faint bands in samples LV92 B- and LV129 B- are arrowed in black.

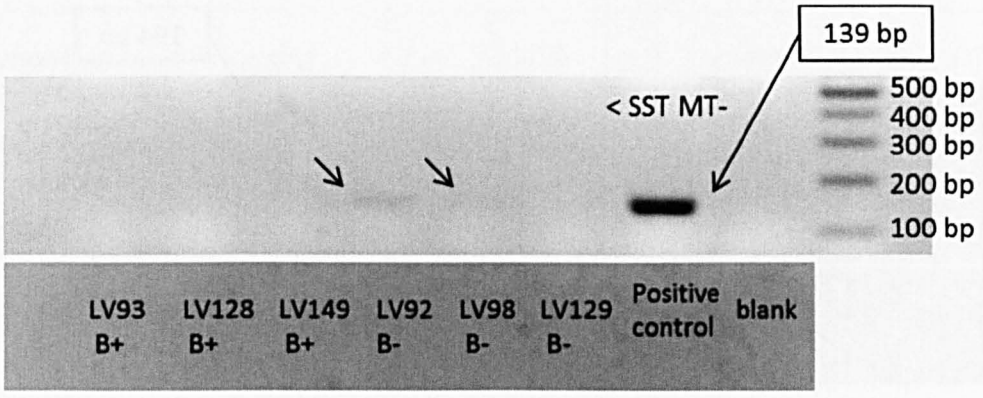


Figure 3.16: RT-PCR for the *MRM2* gene on strains >SST. The positive control is a <SST MT- sample where it is known that the gene was expressed. The faint bands in samples LV92 B- and LV98 B- are arrowed in black.

I can thus conclude that the RNA-seq results (Table 3.6) were mostly validated proving that three of the MT-biased genes of *P. multistriata* were not expressed in sexually immature samples (>SST). The fourth gene (*MRP2*), although expressed in sexually immature samples (>SST), resulted lowly expressed compared to the sexually mature (<SST) and did not show differential expression between opposite MTs. Further analyses in qRT-PCR could be performed to calculate the expression change of the MT-biased genes between <SST and >SST strains.

3.4 Discussion

3.4.1 The 'sensing phase' during sexual reproduction

The results of the analysis of the transcriptome dataset produced at the early phases of sexual reproduction showed that, among the 91 DEG, four out of five MT-biased genes resulted highly expressed at the beginning of sexual reproduction as compared to the control strains grown in mono-culture. Moreover, their expression increased in a time-dependent manner from the early (T1 = 10:30 a.m.) to the late time point (T2 = 02:30 p.m.). *MRP3* was the only, of the five MT-biased genes, to not be involved in the early phase of sexual reproduction showing no expression differences between sexualised strains and control strains.

Observing sexual reproduction in pennate diatoms, from the early stages in which cells of the opposite mating type 'sense' each other, through pairing of gametangia till gametes formation and conjugation, it can be hypothesized that sexualisation is driven by a chemical cue. This cue is represented by sex pheromones, whose presence has been detected (Gillard *et al.* 2013, Moeys *et al.* 2016) or inferred (Sato *et al.* 2011) in two pennate benthic diatoms.

In the araphid pennate *Pseudostaurosira trainorii,* pheromone activity was experimentally documented (Sato *et al.*, 2011) and the authors showed that female (MT-) strains constitutively secret a pheromone ph-1 that induced male sexualisation, i.e. production of gametes, and production of a second pheromone (ph-2). Ph-2 triggered female sexualisation, and these cells probably started producing a third pheromone, ph-3, necessary for male (MT+) gamete motility and attraction. The chemical nature of these sex pheromones has not been characterized yet.

A sex pheromone has been identified in *S. robusta*, where an elaborate multi-step signalling pathway was reported. MT- cells probably produce a primary signal (SIP-) that

activates MT+ cells. MT+ cells start secreting a sex-inducing pheromone (SIP+), responsible of the light-dependent production of L-diproline by MT- gametangia. Both SIPs arrest the cell cycle at the G1 phase. The L-diproline pheromone was capable of attracting MT+ gamentagia (Gillard *et al.*, 2013, Moeys *et al.*, 2016). Gillard et al., (2013) showed that L-diproline production could be detected only after 5 h after illumination in a 12 h time course, proving thus to be strictly light dependent, and that its concentration was exponentially increasing from 5 h up to 10 h, and suddenly decreasing thereafter coinciding with a loss of attraction capacity.

Another example of a microalga in which sex pheromone production was detected is a unicellular Charophycean alga belonging to the *Closterium peracerosum-strogosum-littorale* Complex. It synthesizes two major pheromones involved in early phases of sexual reproduction, where they promote multiple steps all along the conjugation phase. They are known as PR-IP (Protoplast-Release-Inducing Proteins) and PR-IP Inducer. Both are glycoproteins, the first one released constitutively by MT- and inducing the production and release of PR-IP from MT+. PR-IP induces sex cell division with the release of mucilage and gametic protoplast from MT- cells. However, it is still unknown what leads to cell-cell recognition and fusion but the involvement of a third chemotactic pheromone has been hypothesized. The genes encoding the two pheromones are present in both mating types and they resulted differentially expressed (Sekimoto *et al.* 2006, Sekimoto *et al.* 2014).

In all the reported examples a multi-step signalling pathway, mediated by several pheromones, controls mating in different points along the cycle for sexual reproduction.

In these three systems, a primary signal was constitutively produced by one MT to induce the sexualisation process. Its concentration can increase during the process, as for *S. robusta* and *P. trainorii*, and its production can be light-dependent as in *S. robusta* and *Closterium peracerosum-strogosum-littorale.* Consequently, in the latter two species the sexualisation process is light- regulated. On the contrary, in *P. trainorii* successful fertilization was possible in both continuous light and continuous dark conditions. There

are cases, like in the brown alga *Ectocarpus siliculosus*, where the sex pheromones (ectocarpene and homosirene) function as chemo attractant also in other species of the same genus and in other genera. Moreover, their biosynthetic pathways have been reported also from the related stramenopile diatoms with identical fatty acids precursor and with lipoxygenases of identical positional specificity (Frenkel *et al.* 2014). However, in the case of diatoms those metabolites are used for chemical defence (Frenkel *et al.* 2014).

A multi-step sexualisation system, most probably mediated by pheromones, is also present in *P. multistriata.*

The presence of chemical communication in *P. multistriata* is supported i) by flow cytometry data showing an arrest in the G1 phase of cells in chemical contact with the opposite MT during the early phases of sexual reproduction, ii) by the molecular data of the sensing transcriptome, where it was observed that cell cycle arrest in G1 induces expression changes in around 9% of the genes (Basu *et al.* under revision), iii) by the density-dependent mechanism triggering the mating recognition (Scalco *et al.* 2014) and iv) finally by the results reported in this chapter specifically for the five MT-biased genes.

Four out of five MT-biased genes studied with the 'sensing transcriptome', showed direct evidences of their involvement in the early phase of mating recognitions during sexual reproduction. They were differentially expressed in relation to MT in strains below the SST, and were up-regulated in a time-dependent manner, suggesting that they are involved in the sexualisation phase. It is possible to speculate that one or more of the four MT-biased genes is encoding for a signal molecule (cytostatic/chemotactic pheromone) or receptor that activates only when cells become sexually mature, and thus induces or is involved in the multi-step signalling process. The published data on the presence of pheromones in diatoms refer to benthic species, where the production of attracting pheromones makes sense since one cell can move towards the other gliding on a solid substrate or through the action of phylopodia, as in *Pseudostaurosira trainorii.*. *P. multistriata* is a planktonic species that lives in the water column, and specific adaptations

should have evolved in this species to allow encounters between cells of opposite mating type. However, there are several examples of chemical cues acting in planktonic species. Examples are allelopathic compounds that can harm other species (e.g., (Tillmann and John 2002, Paul *et al.* 2009, Lyczkowski and Karp-Boss 2014)), or induce transitions between different life stages (e.g., (Fistarol *et al.* 2004)). It is therefore possible that sex pheromones are active also in planktonic diatoms, provided that a high concentration of cells is reached for a sufficient time to allow its perception by the neighbouring cells. In the case of *Pseudo-nitzschia*, these conditions can be met during a bloom and, indeed, the two reports of massive sexual reproduction of *Pseudo-nitzschia* species in the natural environment have been recorded during a bloom (Holtermann *et al.* 2010, Sarno *et al.* 2010). It has been also suggested that needle-shaped *Pseudo-nitzschia* cells and colonies naturally aggregate in calm conditions of the water column, facilitating encounters between cells (Botte *et al.* 2013). Sexual reproduction is a relatively rapid event and the production of sex pheromones triggered by a density-dependent mechanism, can represent a successful adaptive trait allowing the pairing of complementary gametangia.

3.4.2 Regulation of *Pseudo-nitzschia multistriata* MT-biased genes

The statistical analysis conducted on the expression data of four MT-biased genes along a 24 h time course indicated significant variations only for one gene out of four. The expression of *MRP2*, *MRM1* and *MRM2* was not regulated by the diel light cycling or according to the cell cycle phases. On the contrary, *MRP1* presented a significant point of low expression rate at 10:00 am (T1). Observing the REST analysis of *MRP2*, a low expression rate at 10:00 am (T1) was equally detected, but it was not confirmed by the statistical analysis.

MRP1 is not regulated by light or by cell cycle. Considering the figure reported below (Fig. 3.17), if regulated by light, *MRP1* would have shown a decrease in expression after

8:00 pm (T5) following the light:dark cycle, which has been demonstrated to affect the expression of light-dependent genes (Siaut *et al.* 2007); instead its expression trend kept increasing also at the end of the dark phase. If regulated according to a cell cycle phase, *MRP1* would have shown an expression trend similar to the one exhibited in the histogram for the DNA content (Huysman *et al.* 2010); i.e. if in T1 cells are in G1 for the 80% as in T9, the expression rate of *MRP1* should have been the same at the two time points.

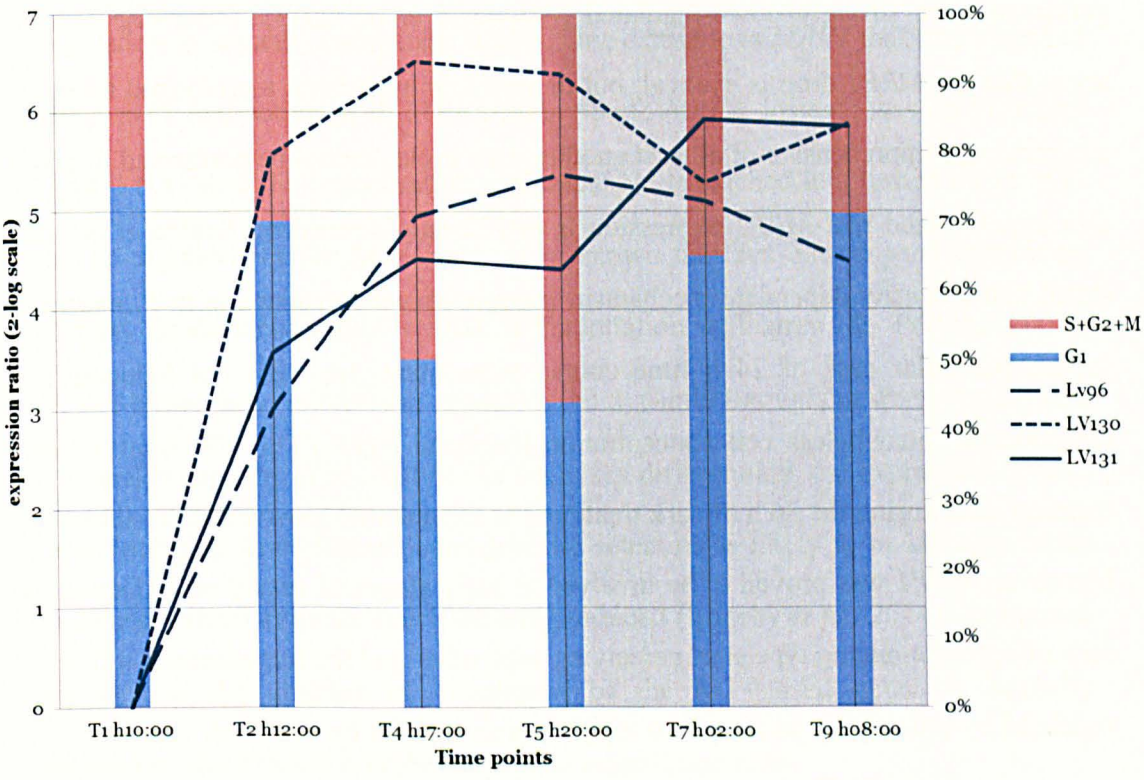


Figure 3.17: Expression trend of *MRP1* in a 24 h light:dark (white background:grey background) cycle. The red and blue bars represent, respectively, the S+G2+M and the G1 cell cycle phases. Line dots represent three MT+ samples.

When organisms are kept in total darkness for extended periods they eventually function with a free-running rhythm. A free-running rhythm takes place when the organism is shielded from any external cue and not adjusted to the natural 24-hour cycle or to any artificial cycle. However in these circumstances, other circadian or ultradian rhythms, such as metabolic, hormonal, etc., become out of phase (Foster and Kreitzman 2004). The low level of expression of *MRP1* at T1, 2 h after re-illumination, could be related to the 36 h dark synchronization treatment. This behaviour was however observed only for *MRP1*,

while the other target and references genes were not affected, indicating that dark synchronization perturbed a specific molecular pathway in which the MT+ specific gene operates. Since *MRP1* encodes for a probable secreted protein, its expression may not be favoured after a prolonged dark condition, possibly because it is energetically too costly. Measures of genome-wide expression of *Thalassiosira pseudonana* during diel growth state transitions showed that after 12 h of darkness genes associated with secretion pathways were strongly down-regulated (Ashworth *et al.* 2013). Otherwise one could argue that the *MRP1* drop is cyclical, but to state this hypothesis a prolonged time course experiment, comprehensive of at least another dark cycle, should be performed. It could be also hypothesized that *MRP1* expression is cell density dependent. Scalco *et al.*, (2014) stated that a density-dependent mechanism triggers sexual reproduction in *P. multistriata*. In this particular case of 24 h time course experiment, we were not inducing sexual reproduction, nonetheless cell concentration increased from T1 to T9, from a low cell concentration during the 36 h of dark treatment to exponential growth after re-illumination. Moreover, *MRP1* was proved to be involved in early stages of sexual reproduction, when cells of opposite mating type start perceiving each other and the hypothesis that it could be regulated by cell-density as needed to trigger sex should not be excluded.

A point that I would like to discuss is the intraspecific variation for patterns of sex-biased gene expression. The potential for sex-biased gene expression to evolve has been discussed by (Ingleby *et al.* 2014), who inferred that it could vary also between different populations of the same species and different environmental conditions. The strain-specific variability observed for MT+ in *MRP1* might be explained by physiological or genetic differences between strains. However, this difference cannot be attributed to the growth conditions (L:D cycle, irradiance, temperature, etc.) to which the strains were exposed since they were identical for all the strains. Differences cannot be attributed to cell size either, because all MT+ strains had similar average cell size. High genetic variability of the MT+ strains

should be also excluded, because all three MT+ strains are siblings. A clear explanation for such strain-specific variability could not be identified. Anyway, very little is known about the modality in which sex-biased gene expression relates to sex-specific fitness and about how sex-biased gene expression and conflict vary throughout development or across different genotypes, populations, and environmental conditions.

The results obtained testing MT-biased genes expression suggest a different mating type-specific regulation in sexually competent strains. One example is *MRP1* that was switched off in strains >SST and activated in MT+ strains <SST; *MRP2,* instead, was on in strains >SST of both MTs, once the sexualisation size threshold is reached it is turned off in MT- strains and up-regulated in the MT+ ones. It is known that sex-biased gene expression becomes most pronounced after sexual differentiation (Ellegren & Parsch, 2007). Moreover, sex-biased gene expression appears to be dynamic throughout development in a number of species (Ingleby *et al.*, 2014). As the sexes differentiate, the expression of sex-biased genes increases since sexually antagonistic selection is likely to be stronger when distinct male and female traits are specified and produced (Ingleby *et al.*, 2014). In the case of *P. multistriata,* the absence of expression of the MT-biased genes in sexually undifferentiated strains (>SST) corroborates these latter statements.

Chapter 4

The challenge to discover mating type locus in *Pseudo-nitzschia multistriata*, a genetic approach: conservation of the MT locus between *Seminavis robusta* and *Pseudo-nitzschia multistriata*, and Bulked Segregant Analysis (BSA) in *P. multistriata*

4.1 Introduction

The transcriptomes of two diatoms, *Pseudo-nitzschia multistriata* and *Seminavis robusta*, have been sequenced within the project "A deep transcriptomic and genomic investigation of diatom life cycle regulation" funded by the Joint Genome Institute (http://genome.jgi.doe.gov/Adeeregulation/Adeeregulation.info.html). The project aim was to sequence the transcriptome of two pennate diatoms with similar life cycle features but distinct ecological niches, planktonic for *P. multistriata* and benthic for *S. robusta*, quite separated in terms of phylogeny (Fig. 4.1), in order to identify genes expressed in different mating types and during distinct phases of the sexual reproduction.

The transciptomic analyses of *P. multistriata* focused mainly on the identification of MT-biased genes (Chapter 2), while the one of *S. robusta* focused on the study of the cell cycle phases and meiotic genes (Patil *et al.*, 2015) (Wim Vyverman personal communication). Vanstechelman et *al.* (2013) provided the first and attempt to identify the MT determining region in *S. robusta*, which is the only information available for diatoms up to now. The Authors constructed a sex-specific linkage map based on AFLP markers to identify the MT determining region. Segregation and linkage analysis of 463 AFLP markers on 116 individuals (57 MT+ and 59 MT-) of an F1 mapping population were analyzed to find markers co-segregating with each mating type. A QTL analysis was further performed to confirm the monogenic nature of mating type. The analysis resulted in the identification of MT+ as the heterogametic sex in *S. robusta*. Three transcripts were identified in the genomic scaffold containing the MT locus, with three domain hits: a leucine-rich repeat receptor-like protein kinase (LRR) (PLN00113); a superfamily of DNA/RNA helicases SF2; and a super family of S-adenosylmethionine-dependent methyltransferases (SAM) (Vanstechelman, 2013). Other conserved domains flanking the MT locus were a protein kinase and a Hedgehog/Interin. Gene structure prediction identified the gene configuration of a SF2-family related Helicase/S-adenosyl methyltransferase (HEL-SAM) as member of

the DNA methyltransferase 5 (DNMT5) protein family. Vanstechelman found also that HEL-SAM had a homolog in *P. multistriata.* Such discovery prompted the sequencing of the HEL-SAM homolog in *P. multistriata* (presented in this Chapter). This analysis was aimed at testing the possible conservation of the sex locus in the two pennate diatoms, studying the polymorphisms pattern.

The pattern resulted not to be different between opposite MTs (see the Results section), and it was thus decided to produce an F1 mapping population of *P. multistriata* (see Chapter1, Fig. 1.6) and to perform a Bulked Segregant Analysis (BSA). BSA is a quantitative trait loci (QTL) mapping technique for identifying genomic regions containing loci affecting the trait of interest (Magwene *et al.*, 2011), in this case the MT locus. This method relies on the co-segregation of unknown trait loci and genetic markers with known chromosomal locations (Claesen *et al.*, 2013). Starting with a segregating population from a genetic cross, individuals are assayed for the focal trait (in this case, the different mating type) and two pools (bulks) of segregants are created. Genotype frequencies are estimated for the two bulks, based on the marker frequencies observed in the pooled DNA samples. Allele frequencies of the two bulks are expected to be approximately equal in genomic regions without loci affecting the trait (Magwene *et al.*, 2011), but they should differ at genomic regions containing the MT locus (loci). The advent of next generation sequencing (NGS) allows a fast identification of a huge number of single nucleotide polymorphisms (SNPs) on a genome-wide scale, providing a very dense set of markers. When combined with segregant pooling, NGS-BSA allows for simultaneous SNP-discovery and mapping of trait loci throughout the entire genome (Claesen *et al.*, 2013). Hence, the BSA-sequencing approach allows for detecting markers in linkage with causal loci as well as allelic biases at the causal loci themselves. Furthermore, sequencing data yields counts of alleles at polymorphic loci and thus provides a simple and intuitive way of estimating allele frequencies (Magwene *et al.*, 2011).

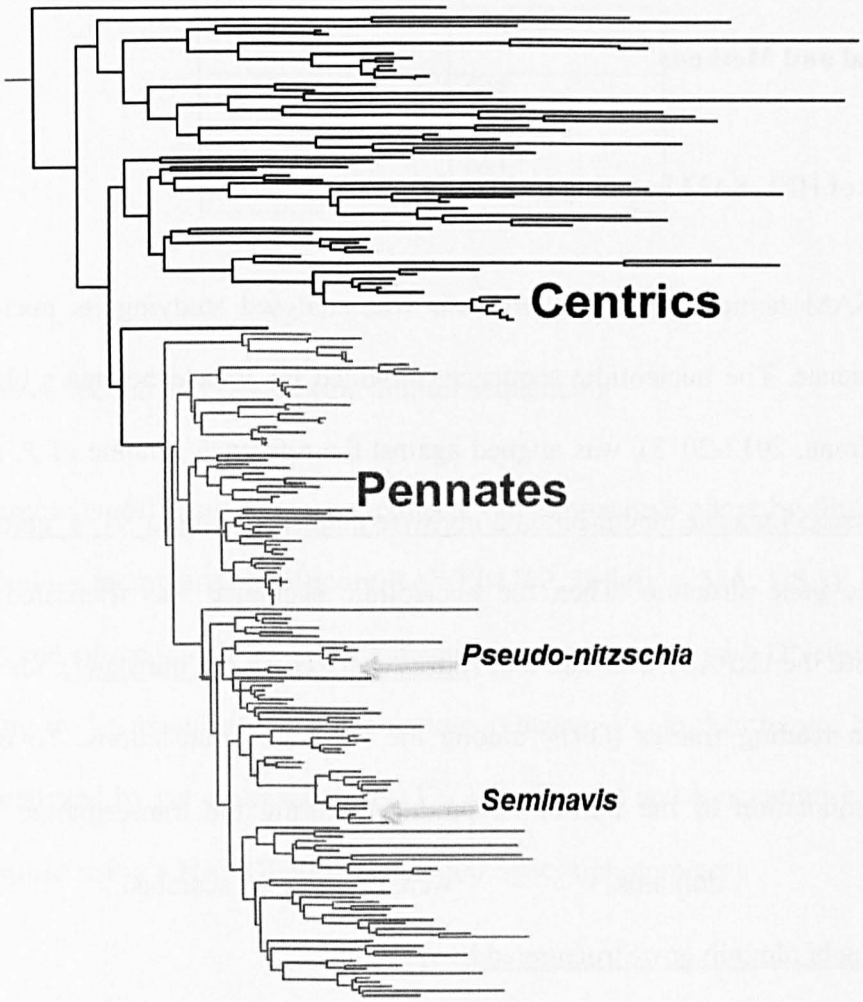


Figure 4.1: Phylogenetic tree built with 18S rDNA of diatoms (Kooistra *et al.*, 2003), the position of the two genera of interest is highlighted.

4.2 Material and Methods

4.2.1 Study of HEL-SAM homolog in *P. multistriata*

The HEL-SAM homolog in *P. multistriata* was analysed studying its nucleotidic and proteic sequence. The nucleotidic sequence, provided by Vanstechelman's blast analysis (Vanstechelman, 2012-2013), was aligned against the reference genome of *P. multistriata* http://gbrowse255.tgac.ac.uk/cgi-bin/gb2/gbrowse/maplesod_psnmu_v1_4_gbrowse255/ to visualize the gene structure. Then the nucleotidic sequence was translated to protein sequence with the ExPASy translate tool (http://web.expasy.org/translate/), identifying the correct open reading frames (ORF) among the six frame translations. To confirm the functional annotation of the transcripts produced during the transcriptome annotation, conserved domains were searched through http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi.

4.2.2 Cultures for sequencing

The strains of *P. multistriata* used for sequencing HEL-SAM homolog were one MT+ and one MT- (Table 2.2). Strain B856 was the one used for sequencing the *P. multistriata* genome, while B857 is a sibling. Both strains have been used for RNA-seq (see Chapter 1, Fig.1.6, Chapter 2, Table 2.1). Protocols for f/2 culture medium (Guillard, 1975) preparation and for growth condition are the same as those reported in Chapter 2.2.7.

Table 4. 1: Strains of *Pseudo-nitzschia multistriata* used for sequencing of HEL-SAM homolog. Reported the strain code, the mating type, and the DNA extraction date.

Strain code	Mating type

	(Mt)
B857	MT-
B856	MT+

4.2.3 Sample collection and DNA extraction for sequencing

Pseudo-nitzschia multistriata cells were collected in exponential phase by filtration on 1.2 µm nitrocellulose membranes (Millipore RAWP04700, Billerica, MA, USA). Filters were flash frozen and stored at -20 °C. DNA extraction was performed with DNeasy Plant Mini Kit according to the manufacturer's instructions (Qiagen, Venlo, Limburgo, Nederlands). DNA was analyzed by gel electrophoresis (1% agarose w/v) and concentration and quality were determined using a NANODROP (ND 1000 Spectrophotometer).

4.2.4 Primer design, PCR, purification and sequencing of HEL-SAM homolog in *P. multistriata*

The protocol for primer design was reported in Chapter 2.2.6. To cover the full length of the gene (6893 bp), 13 primer pairs were manually designed on the transcript sequence of the gene homolog to HEL-SAM (ID: 0081690.1). PCR amplifications were conducted on genomic DNA of MT+ and MT- strains. PCR reactions were carried out in a volume of 100 µl: gDNA 2.5 µL, oligo fw (2.5 µM), oligo rv (2.5 µM), PCR reaction buffer with $MgCl_2$ 10X (Roche, Basel, Switzerland), dNTP (2 mM), Taq DNA Polymerase (0.25 U/µL) (Roche, Basel, Switzerland). The thermal profile of amplification varied depending on the fragment to be amplified. The products were checked on 1 % agarose gel in TAE buffer and ethidium bromide staining with a 1 Kb ladder, to recognize the size of the band amplified (Gene Ruler 1 kb DNA Ladder - Thermo Scientific Fermentas, Waltham, Massachusetts, USA). The PCR products were purified with QIAquick PCR purification

kit (Qiagen, Venlo, Limburgo, Nederlands) according to the manufacturer's instructions. The sample for the sequencing reaction was composed by purified DNA [15 fmol/µl] + primer [4.5 pmol/µl] in a final volume of 20 µl. Sequence reactions were obtained with the BigDye Terminator Cycle Sequencing technology (Applied Biosystems, Foster City, CA), purified in automation using the Agencourt CleanSEQ Dye terminator removal Kit (Agencourt Bioscience Corporation, 500 Cummins Center, Suite 2450, Beverly MA 01915 - USA) and a robotic station Biomek FX (Beckman Coulter, Fullerton, CA). Products were analyzed on an Automated Capillary Electrophoresis Sequencer 3730 DNA Analyzer (Applied Biosystems).

Table 4.2: List of primer pairs used for PCR sequencing of gene 0081690.1 in *P. multistriata*; primer position along the gene, primer name, sequences and amplicon size are reported. External to the transcript means that the primer was designed in the external genomic region flanking the transcript.

Primer position along gene 0081690.1	Primer name	Sequences	Amplicon size
69-90	3.59 F1	5'- GCACCAACCTGTATCTGTTTTC -3'	671 bp
723-739	3.59 R1	5'- CGCAAATCTGCACCGTC -3'	
667-686	3.59 F2	5'- GCGAGGGTGATGTGCTCTAT-3'	675 bp
1322-1341	3.59 R2	5'- CTACAATACCATGCGTCGGG -3'	
1275-1292	3.59 F3	5'- CAGAAATGGCCCGAGAAG -3'	667 bp
1922-1941	3.59 R3	5'- GGCCCTGGATATACTCTTGC -3'	
1842-1860	3.59 F4	5'- GCCGCTGTGGAATTTCTTG -3'	682 bp
2507-2523	3.59 R4	5'- GTTCCTTTGCACGGTCG -3'	
2443-2461	3.59 F5	5'- CCCATGATCGGATTCGTTC -3'	680 bp
3123-3142	3.59 R5	5'- GGCAAAGTCGTGCTTTTGAC -3'	
3064-3082	3.59 F6	5'- CCAACTTCGAAAACAACGC -3'	672 bp
3717-3735	3.59 R6	5'- CGAAGAGGGTTCTCGTTCC -3'	
3646-3665	3.59 F7	5'- CGGGAGAAATAGCTCTCTCG -3'	704 bp
4333-4349	3.59 R7	5'- GACGGCCCTGTTGGATG -3'	

4271-4292	3.59 F8	5'- CTAGACTGGATTAATGCCCCTG -3'	670 bp
4921-4940	3.59 R8	5'- CTTTTGTATCCGGCTCTCCC -3'	
4025-4045	3.59 F8'	5'- GGTCGCATCAATGGAATCTAC-3'	631 bp
4636-4655	3.59 R8'	5'- GTGGAACATGGTTCTTGCAG-3'	
4866-4885	3.59 F9	5'- GCGAAGCAACCGACTGTATC -3'	699 bp
5546-5564	3.59 R9	5'- GGCCTTCGATAGATGGAGC -3'	
5488-5504	3.59 F10	5'- GAAGAAGGCAAACGCCG -3'	667 bp
6135-6154	3.59 R10	5'- GACCTCGAGCTTCTCACCAC -3'	
6077-6095	3.59 F11	5'- CGCTGTGCTGTTCAAGAGG -3'	717 bp
6775-6793	3.59 R11	5'- GCGTTCCATAATCCAGTCG -3'	
6683-6702	3.59 F12	5'- GCCGTAGGAAGGGTATTTCG-3'	710 bp
External to the transcript	3.59 R12	5'- GTACGAGTAACTCGCAGTATCACAC-3'	

4.2.5 Editing and sequence alignment of HEL-SAM homolog

The sequences produced were edited with Chromas Lite 2.01 (http://technelysium.com.au/) paying particularly attention for SNPs in the chromatogram. The sequences were mapped against the reference gene sequence, downloaded from the *P. multistriata* genome browser, with LASTZ sequence alignment program through Galaxy, an open source web-based platform (Giardine *et al.*, 2005, Kobiyama *et al.*, 2007, Bertrand *et al.*, 2012), and visualized with Tablet 1.14.04.10 (https://ics.hutton.ac.uk/tablet). Further analysis to detect genes heterozygosity was performed with the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011, Thorvaldsdóttir *et al.*, 2013) applied to all the genomic data sets that became recently available for P. multistriata (Table 1.3). The data sets comprised 22 RNA-seq libraries (11 MT+ and 11 MT-) and 6 genome sequencing (4 MT+ and 2 MT-).

4.2.6 Production of an F1 mapping population for BSA

A full-siblings (FS) family was produced from crosses between strains B854 MT+ and MVR1041.4 MT-. Isolation of F1 initial cells was carried out manually by single-cell isolation with a micropipette (Andersen, 2005). F1 initial cells were transferred to 24-well culture plates containing 2 ml f/2 medium. Once the cultures reached a good concentration, i.e. when the bottom of the well was covered by *P. multistriata* chains, they were transferred to 25 cm^3 flasks filled with 20 ml of f/2 medium. Strains were incubated at 20°C and 130 µmol photons $m^{-2}s^{-1}$, provided by cool white fluorescent tubes TLD 36W/950 (Philips, Amsterdam, Nederland) and natural light, to speed up growth. After almost four months of weekly transfers, the F1 cultures reached the sexualisation size threshold (cell length < 60 µm). The smallest cells of each culture were re-isolated following the procedure described above, with the aim of creating cultures of uniform size.

4.2.7 Mating type determination of the F1 progeny

The mating type of the F1 progeny was determined by crossing each individual strain with two MT+ and two MT- strains, already tested to be good reference couples: SH20, MT- and MVR171.8, MT+, MVR171.1 MT- and B856, MT+. For the method of mating type tests, please refer to Chapter 2.2.8.

4.2.8 Sample collection, DNA extraction and bulks preparation for BSA

30 MT+ and 30 MT- segregants were selected to perform BSA. Cultures were made axenic growing them for 4-5 days in f/2 medium supplemented with three antibiotics: Streptomycin (0.1mg/ml), Penicillin (0.1mg/ml) and Ampicillin (0.5mg/ml). Bacterial contamination of the cultures was checked under epifluorescence microscope by DAPI

staining. *P. multistriata* cells were collected in exponential phase (~ 200,000 cell·ml^{-1}) by filtering 200 ml of axenic culture on 47 mm 1.2 µm nitrocellulose membranes (Millipore RAWP04700, Billerica, MA, USA). Cell growth was monitored by estimating cell concentration using a Sedgewick-Rafter counting chamber. The algal pellet was collected from the filter and frozen at -20°C. The DNA was extracted following a Phenol-Chloroform extraction method (Vanstechelman *et al.*, 2013) with slight modifications that include cell disruption by adding 400 mg of 0.2-0.3 mm diameter silica beads and vortex mixing at 30 hertz for 85 seconds (3 times), cooling the pellet on ice between the vortex mixing. The extracted DNA was ethanol precipitated, air dried, dissolved in 50 ul of sterile water and stored at -20 ^{0}C until sequencing. DNA was analysed by gel electrophoresis (0, 8% agarose w/v) and with a Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) to assess concentration. Subsequently a MT+ and MT- bulk was constructed by pooling the same DNA amount for each of the 30 segregants with the corresponding MT.

4.3 Results

4.3.1 Testing HEL-SAM as MT locus in *P. multistriata*

In *Seminavis robusta,* preliminary analyses detected three genes to be part of the MT-locus, of which only the HEL-SAM has been found in the genome of *Pseudo-nitzschia multistriata* (Fig. 4.2). The *P. multistriata* gene model PSNMU-V1.4_AUG-EV-PASAV3_0081690.1 was found to be the homolog of *S. robusta* HEL-SAM. It was located on PsnmuV1.4_scaffold_4-size_463035:377366..384685 (+ strand) and contained five introns. Two transcripts were overlapping the gene model: comp21438_c0_seq2.1, in position 377322..384165 (+ strand), and comp32838_c0_seq1.1, in position 384166..384696 (+ strand), for a total length of 7375 bp. However, the transcripts were incorrectly assembled as both presented incomplete ORF with no stop codon. Moreover the RNA-seq reads further confirmed the truthfulness of the gene model prediction against the transcripts structure. Since the transcripts follow one the other without interruption (377322..384165 -384166..384696) the sequence was manually corrected merging the two transcripts in one and resulting in a sequence of 6893 bp with complete ORF (RF+1) of 6588 bp and with a protein sequence of 2195 AA (equally to the protein predicted to be encoded by the gene model) possessing the two conserved domains of AdoMet_MTases super family (S-adenosylmethionine-dependent methyltransferases) (SAM) and HepA (Superfamily II DNA or RNA helicase, SNF2 family) (HEL) (Fig. 4.3). The 13 primer pairs were designed to cover the full length of the sequence (Fig. 4.2). The amplicon size was of 600-700 bp according to the sequencing station capacity.

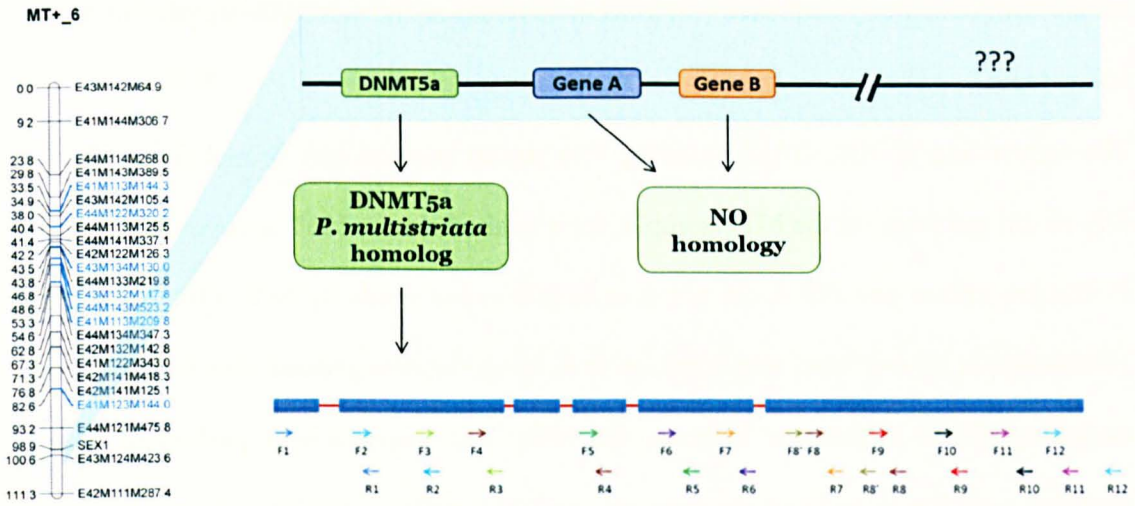


Figure 4.2:MT-locus of *S. robusta* showing on the left the linkage map where the locus was indentified. (Vanstechelman *et al.*, 2013) and the genes detected in the MT-locus. On the bottom the gene structure of HEL-SAM homolog in *P. multistriata* and the positions of the 13 primer pairs designed on it.

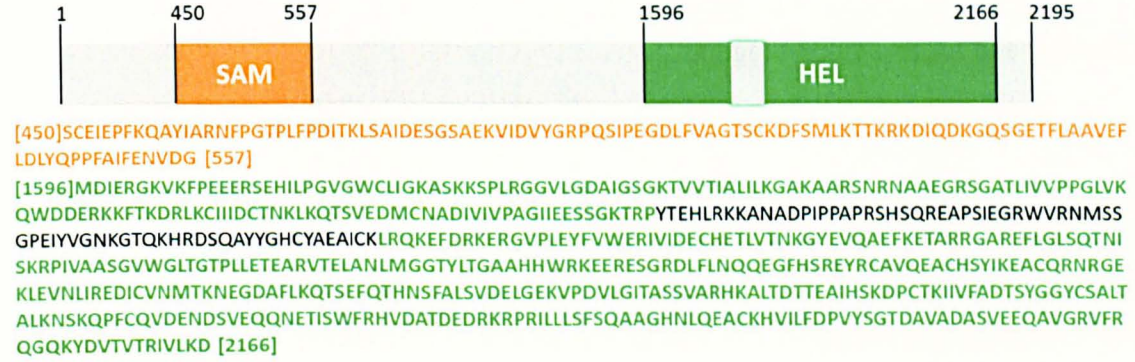


Figure 4.3: Scheme showing the 2195 AA protein codified by HEL-SAM homolog in *P. multistriata*. The HAdoMet_MTases super family (SAM) domain (orange) and the HepA Superfamily II DNA or RNA helicase, SNF2 family (HEL) domain (green) that has a gap from position 1742 to 1810.

The aim of the sequencing was to test the presence of a MT-related pattern of Single Nucleotide Polymorphisms (SNPs) to validate the hypothesis of a conserved sex locus between *Seminavis robusta* and *Pseudo-nitzschia multistriata*. However, no SNP peaks were detected when analysing each sequence. The alignment of all the forward and reverse sequences of the fragments of HEL-SAM homolog of both the MTs against the reference scaffold did not provide evidence of nucleotide variation between the MT+ and MT- samples. The only variations observed were due to sequencing errors (N), not confirmed

on the complementary sequence (forward or reverse) of the same primer pair on the same MT sample.

The sequencing of HEL-SAM homolog was partial because two primer pairs F/R 8 and F/R 10 did not work on the MT+ sample, for a total of almost 1337 uncovered bases. In the PCR runs, primer pair F/R 8 did not give amplification result for MT+ (B856), while F/R 10 resulted in double band amplification that, however, was present also in MT- samples, excluding allelic differences between the MTs. The analysis was performed when no genomic tools for *P. multistriata* were yet available. SNPs variability was totally screened by IGV to visualise those gaps resulted by the sequencing procedure. A total of 15 MT+ and 13 MT- sequences were aligned against the reference genome (belonging to a MT+) but no polymorphisms in heterozygosis according to MT were observed for the HEL-SAM homolog and for its flanking regions.

4.3.2 Production of an F1 mapping population

A F1 mapping population of 152 strains was produced by crossing two parental strains of complementary mating type. However, some of the strains died over the time required to reach the cell size threshold for sexualisation (about for months) and the mating type of only part of the F1 progeny could be determined unambiguously, yielding a mapping population of 41 MT + and 52 MT- strains. 30 MT+ and 30 MT- segregants were selected to construct one MT+ and one MT- bulk. DNA for each bulk will be sent to our sequencing provider who will perform library preparation and Illumina sequencing following the standard protocols.

A first trial of BSA sequencing was already carried out at the beginning of 2015, but the low quality of the sequencing, due to a bacterial contamination, and the insufficient coverage of MT+ and MT- bulks made it impossible to be analysed. The protocol for bulks

preparation was thus improved, adding antibiotic treatment of the cultures and collection of the latter by filtration.

4.4 Discussion

One of the aims of my PhD project was to test if the mating type of *P. multistriata* was genetically determined. The assumption was that *P. multistriata* MT-locus follows the sex-determining mechanism in which MT+ is heterogametic and MT- is homogametic. This assumption was proved to be correct by the results of MT distribution in the F1 progeny produced by sexual events between two strains of opposite mating type resulting in a sex ratio of almost 50:50.

The attribution of the MT to a very high number of F1 strains obtained from a single cross of two parental strains of complementary mating types, carried out with the aim of building an F1 mapping population, showed that mating type ratio in *P. multistriata* is balanced: 41 MT+ and 52 MT-. This is a proof that sex determination in *P. multistriata* is genetically determined and that the MT locus should be heterozygous for one of the MT. The law of segregation of Mendel states that every individual contains two alleles for each trait, and that these alleles segregate during meiosis. Thus, each parent contributes with a single allele copy to their offspring.

To the best of my knowledge, there is only one report on mating type ratios in diatoms: this is a study carried out on a natural population of the pennate benthic diatom *Nitzschia longissima* (Davidovich *et al.*, 2006). The Authors assessed the mating type of 68 clonal cultures isolated from the coastal habitats near the Karadag Biological Station (Crimea, Ukraine) resulting in 35 "male" (MT+) and 32 "female" (MT-) clones. The balanced sex ratio provides a further support to the fact that in pennate heterothallic diatoms sex is genetically determined. Davidovich *et al.* (2006) observed that 21 of 35 MT+ clones were capable of intraclonal sexual reproduction (facultative andromixis) and that their progeny consisted of both MT+ and MT- in a balanced ratio. These observations suggest that the MT+ is the heterogametic sex in this species.

The almost 50:50 sex ratio of the progeny obtained from the cross of *P. multistriata* performed in the laboratory find support also in the data obtained from field populations that were tested for mating type attribution on a wide number of strains isolated during the bloom season of this species in different years: 2008, 2009 and 2010, (Scalco, 2013; for details of the method see Chapter 2). Interestingly, the percentages were relatively balanced for strains isolated in 2009 (50.8% MT+ and 49.2% MT-) and 2010 (37.9% MT+ and 56.1% MT-), but in 2008 the 92.2% of the strains turned out to belong to MT-. These very puzzling results raise questions about the mechanisms that determine mating types in this diatom. If sexes/mating types are determined by the presence or absence of an allele on a single gene locus, the random segregation of genes at meiosis will produce a balanced sex ratio.

Bull (1983) stated that sex ratio selection is the underlying force shaping the evolution of sex determining systems. The Author proposed that transient linkage disequilibrium between sex determining alleles and genes under strong positive selection could destabilize sex determination by causing distorted sex ratios in a population. Unbalanced sex ratios have been reported in some organisms, e.g. lizards, as the result of the interplay between genotypic sex determination and environmental sex determination (Uller *et al.*, 2007). In the brown algae *Laminaria saccharina* and *L. religiosa* it has been shown that sex ratio can be modified by environmental stressors (Bartsch *et al.*, 2008).

Werren and Beukeboom (1998) proposed that the sex determining system consists of parental sex ratio genes, parental effect sex determiners and zygotic sex determiners, which are subject to different selection pressures due to differences in their modes of inheritance and expression. The Authors reviewed the role of genetic conflict as the driving force to explain the evolution of several sex determining mechanisms. Genetic conflict occurs when different genetic elements within a genome are selected to "push" a phenotype in different directions, providing the trigger for evolutionary changes in sex determination. These theories can be the starting point to investigate the sex determining mechanisms in

P. multistriata, merging together the knowhow acquired from laboratory and field observations.

I searched for heterozygosity at SNP level for the *P. multistriata* homolog of HEL-SAM, the putative mating type determining gene in *S. robusta*. The negative result suggests that the structure of the sex-locus is not conserved between the two species, and that differences between the two might be substantial. Both species are pennate raphid diatoms but cluster in different clades in diatom phylogenies built with 18S (Kooistra *et al.*, 2003) (Fig. 4.1). They also have different habits, benthic and planktonic, respectively, and differences are also recorded in the behaviour during the sexual phase. In *S. robusta,* MT+ cells swim actively towards MT- ones (Gillard *et al.*, 2013) while there is no clear evidence of cell attraction in *P. multistriata* (Scalco *et al.*, 2015).

As illustrated in the Introduction of this thesis (Chapter 1), a considerable diversity of sex determination systems is present amongst eukaryotes. Dual sex chromosome systems, in which either the female (ZW/ZZ) or the male (XX/XY) is heterogametic, are common in vertebrates and plants. Other systems, as in *Drosophila melanogaster* and *Caenorhabditis elegans*, are set by the ratio of the number of X chromosomes to sets of autosomes (X:A) (Haag & Doty, 2005). In some macroalgae and bryophytes there is a haploid phase determination system (UV system) (Bachtrog *et al.*, 2011). In contrast to animals and plants, fungal and algal cell-type identity is orchestrated by a more restricted chromosomal region, known as the mating type (MAT) locus. In fungi mating types occur in two general patterns: i) bipolar, as single genetic locus occurring in two alternative forms (a or α) of a unique gene (Metin *et al.*, 2010); ii) tetrapolar, where two unlinked genomic regions establish cell identity, one locus encoding pheromones and pheromone receptors, the second encoding homeodomain transcription factors (Fraser *et al.*, 2004).

Remarkable diversity of independently evolved sex-determining mechanisms exists even in closely related lineages. Teleosts fishes, for example, are characterized by sex-

determining mechanisms that range from those using environmental cues to those genetically controlled (Star *et al.*, 2016). Furthermore a wide variety of master sex determining genes has been described in different genera, i.e. *dmY*, *gsdfY* and *sox3Y* in the genus medaka, *amhr2* in fugu, *amhy* in Patagonian pejerry and Nile tilapia, *dmrt1* in half-smooth tongue sole, *gdf6Y* in killifish and *sdY* in rainbow trout (Star *et al.*, 2016).

The absence of conservation between *S. robusta* MT-locus and *P. multistriata* is a further proof that considerable difference exist in the mechanisms that regulate the mating type of these two model diatom species.

Chapter 5

General conclusion and future perspectives

My PhD project provided new insights into the molecular mechanisms related to the mating type determination system of the marine planktonic diatom *Pseudo-nitzschia multistriata.*

A differential expression analysis of the genes of opposite mating types through a transcriptomic approach and a subsequent validation of the results in qRT-PCR resulted in the identification of five MT-biased genes, three MT+ related (*MRP1, MRP3* and *MRP3*) and two MT- related (*MRM1* and *MRM2*) (Chapter 2). These genes were expressed during the vegetative phase in monocultures below the sexualisation size threshold (SST), i.e. when cells were sexually competent, thus proving evidence for the mating type-specific expression of the five genes. The expression of the five genes was analysed also in the early phases of mating type recognition, in an experiment in which the opposite mating types were kept physically separated but were allowed to exchange chemical signals through the free flux of the culture medium. Four out of the five genes showed considerably higher expression in the sexualized samples, i.e. the strains in 'chemical contact', as compared to the monocultures of parental strains. Moreover, gene expression increased in relation to time, being higher after six hours from the beginning of the experiment. These results demonstrated the unequivocal involvement of the four mating type-related genes in the sensing mechanism between cells of opposite mating type during the sexual phase.

The results of further experiments aimed at studying the regulation and functional role of the four MT-biased genes that showed a clear involvement in the sexual phase are presented in Chapter 3. A 24 hour time course experiment (12L:12D) including the analysis of expression fold change by qRT-PCR on three pairs of strains was carried out to test a possible regulation by light and/or cell cycle phase. The working hypothesis stemmed from literature data, including a publication on the raphid pennate diatom *Seminavis robusta,* showing that the expression of genes involved in the sexualisation

process can be light dependent (Gillard *et al.* 2013, Sekimoto *et al.* 2014). I could not find evidence for the regulation of the MT-biased genes in relation to light or to cell cycle in *P. multistriata*, with the exception of the down-regulation in the expression of *MRP1* at 10:00 a.m., 2 hrs after re-illumination. Further experiments demonstrated that these genes were not expressed in in large-sized strains above the SST, thus further proving that they regulate specific pathways activated only after the reach of the cell size threshold for sex.

Hypotheses on the functional role of the five MT-biased genes were formulated based on a computational characterization. *MRP1* has unknown annotated function but its protein contains a signal peptide suggesting that the protein is secreted; this hypothesis is further confirmed by the prediction of extracellular localization. *MRP2* and *MRM2,* whose annotation was manually revised as probable leucine-rich repeat containing protein, possess a transmembrane region indicating that the proteins likely work as receptors on the cell membrane or on the membrane of an organelle. *MRM1* was annotated as heat shock factor protein 3 with DNA-binding properties indicative of its role in the regulation of other genes as transcription factor. *MRP3* had unknown annotated function.

The sex determination system in diatoms has been studied, only in *Seminavis robusta*, for which the first attempt to identify the MT determining region has been carried out (Vanstechelman *et al.* 2013). It has been possible to conduct a comparative analysis between the two species since one of the genes part of the MT-locus of *S. robusta* (HEL-SAM) had a homolog in *P. multistriata* (Chapter 4). The sequencing of the homolog was performed in search of a MT-related pattern of Single Nucleotide Polymorphisms (SNPs) to validate the hypothesis of a conserved sex locus between *S. robusta* and *P. multistriata.* The negative results suggest that the structure of the sex-locus is not conserved between the two species and that differences between the two might be substantial, also considering

that none of the five MT-biased genes identified in *P. multistriata* were detected in the genome of *S. robusta*.

A proof of the fact that mating type should be genetically determined in *P. multistriata* derives from the assessment of the mating type carried out on a large number of F1 strains produced by a single cross, which provided an almost balanced ratio: 41 MT+ and 52 MT-.

Merging all the information I have obtained on the five MT-biased genes, I can conclude that four of them are involved in signalling processes, and that all of them are likely activated by a primary mating type-determining gene that triggers a cascade of processes in concomitance with the switch between >SST and <SST, leading to stable expression/repression of the genes expressed by one of the two MT.

An hypothetical model of the molecular mechanism at the basis of sexual reproduction in *P. multistriata* is illustrated in the following. Strains above the SST have the MT-biased genes totally switched off or expressed at extremely low levels; *MRP2* is the only gene that was expressed in large cells of both MTs. As soon as strains reach the SST, the mating type is defined, but the primary MT determining gene/s that induces sexual differentiation is still unknown. The MT-biased genes activate and express in a MT-specific manner. MT+ strains express three genes, *MRP1, MRP2* and *MRP3,* at higher levels in respect to the MT- strains during their vegetative growth, while MT- strains express two genes, *MRM1* and *MRM2,* at higher levels in respect to the MT+ strains. However, not all the five genes show the same expression level. *MRP1* could act as primary signalling molecule, which can be present also before the start of the sexualisation phase, activating the entire process at the right moment (Frenkel *et al.* 2014). *MRM2* could act as a receptor of an external cue, possibly the product of *MRP1*. When the two mating types get in contact, the expression level of four of the five genes drastically increases, probably activating the machinery of cell-cell recognition and attraction. It can be hypothesized that MT- cells, upon perceiving the primary signal coming from the MT+, start the transcription of the

gene for MT- pheromone production, possibly mediated by the *MRM1* transcription factor. Consequently, an increase of *MRP2* is induced in MT+ cells, which can produce receptor-like proteins located on the cellular membrane or on the membrane of an organelle. *MRP2* and *MRM1,* showed low expression during the vegetative phase but they increased considerably the expression levels during the sexualisation phase, suggesting that they are regulated by the mating machinery. However, there are several potential mechanisms that might regulate sex-biased gene expression, including alternative splicing of a key gene or involvement of micro-RNAs. These miRNAs target mRNAs with complementary sequences and bind to them to regulate their expression, or to prevent their translation or to destroy them (Ingleby *et al.* 2014).

Future perspectives

Future research should focus on genetic transformation and functional studies to decode the proper functions of these highly differentially expressed MT-biased genes. Sabatino *et al.*, (2015) achieved the first genetic transformation of the planktonic diatoms *P. arenysensis* and *P. multistriata* with the biolistic method, using the H4 gene promoter from *P. multistriata* to drive expression of exogenous genes. In Ferrante's lab (M. Ferrante personal communication) the transformation of the five MT-biased genes is in progress. Four out of five genes have already been cloned upstream of the GFP (green fluorescent protein) to produce a fluorescent fusion protein. Fluorescent tagging will reveal the exact cellular localization of the proteins and will help to better define their roles in the process of sexual reproduction. Moreover, novel tools to modulate gene expression, like overexpression and gene silencing, have been developed for the model species *Phaedactylum tricornutum* and *Thalassiosira pseudonana* (Siaut *et al.* 2007, De Riso *et al.* 2009, Scalco 2013) and are in development also for *P. arenysensis* and *P. multistriata* (Sabatino *et al.* 2015). In case the overexpression of the mating type-biased genes in *P. multistriata* could not be performed, a possible alternative could be the transformation of

MRP1 in bacteria, i.e. *Escherichia coli* (Chen *et al.* 2015). The proteobacteria are a major group of gram-negative bacteria that include a wide variety of pathogens, such as *E. coli.* The choice of proteobacteria has to be connected to the abundance of these organisms found in the cultures of *P. multistriata* (data not shown), so that co-culturing growth condition could be compatible. Therefore we could easily build a bioassay to study the effect of *MRP1* overexpression in MT- cultures. In the best hypothesis, we could observe a direct 'phenotypic' effect in MT- cells, such as gametes production.

A comparative approach on the regulatory regions upstream the MT-biased genes could also be carried out to detect a conserved promoter among the five (Russo *et al.* 2015). Such a result would further confirm their involvement in a regulative cascade activated by a putative primary sex-determining gene and would allow the identification of other downstream regulated genes. The analysis could be extended also to other congeneric species to test the conservation pattern of the overall regulatory pathway. In fact, four out of the five MT-biased genes have been recovered in the genome of *P. multiseries* and in *Fragilariopsis cylindrus*, thus showing that these genes are conserved, at least at the level of phylogenetically closely related species (see Chapter 2).

In the near future, we are planning to improve the transcriptomic dataset on which the analysis illustrated in Chapter 2 was conducted. Having now available RNA-seq data for eight unrelated strains below the SST in the vegetative phase (4MT+ and 4MT-) obtained by the two *P. multistriata* transcriptome projects (the JGI Mating type project and the 'Sensing transcriptome' produced within the PhD project of S. Patil), it could be worth merging the datasets and repeating the differential expression analysis to achieve a more robust result.

I hypothesized that *P. multistriata* MT-locus follows the sex-determining gene model in which MT+ is heterogametic and MT- is homogametic or *vice versa.* For this reason, it

was decided to perform a BSA to search for alleles at polymorphic loci and to estimate allele frequencies. This analysis will permit to assess the heterozygosity of one of the MT. Five MT-biased genes have been already identified in this PhD thesis and more markers linked to the MT-locus and differing between MT+ and MT- will be provided by the BSA analysis. The MT-locus and these additional markers could be used to distinguish the two MTs also in environmental samples (Chen *et al.* 2015) enabling a detailed study on the population dynamics of *P. multistriata*. The LTER (Long Term Ecological Research) station Mare Chiara in the Gulf of Naples is regularly sampled for plankton and the main physical-chemical parameters on a weekly basis since 1984 and samples of environmental DNA have been collected since 2009. The identification of markers linked to the MTs will thus allow a study on MT distribution at sea.

Future analysis to detect genes heterozygosity will make use of the Integrative Genomics Viewer (IGV) applied to all the genomic data sets now available for *P. multistriata*, with no need to use Sanger sequencing to look for SNPs. IGV is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, aligned sequence reads, mutations, copy number, RNAi screens, gene expression, methylation, and genomic annotations. It possesses a VCF mode that stands for Variant Call Format, and it is used to encode SNPs and other structural genetic variants (Robinson *et al.*, 2011, Thorvaldsdóttir *et al.*, 2013).

To detect the differences between the two genomes (MT+ and MT-), and thus detect the loci at which they diverge, the genome re-sequencing of a number of strains of opposite mating type was taken into account. This approach, associated with genetic analysis on the sex determining region, has been proven successful in various organisms (e.g. (Palaiokostas *et al.* 2013, Zhang *et al.* 2015). In a current project the genome re-sequencing of three MT+ and two MT- strains of *P. multistriata* has been recently completed. The

analysis of the re-sequenced genomes with targeted SNPs selection, together with the results of BSA, should allow to identify the MT-locus of *P. multistriata*. Moreover, it will further improve the genome assembly of which I contributed refining the gene models prediction by manual check.

Bibliography

Abane, R. & Mezger, V. 2010. Roles of heat shock factors in gametogenesis and development. *FEBS J.* 277:4150-72.

Adelfi, M. G., Borra, M., Sanges, R., Montresor, M., Fontana, A. & Ferrante, M. I. 2014. Selection and validation of reference genes for qPCR analysis in the pennate diatoms *Pseudo-nitzschia multistriata* and *P. arenysensis*. *J. Exp. Mar. Biol. Ecol.* 451:74-81.

Ahmed, S., Cock, J. M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A. F., Dittami, S. M. & Corre, E. 2014. A haploid system of sex determination in the brown alga *Ectocarpus* sp. *Curr. Biol.* 24:1945-57.

Amato, A., Orsini, L., D'Alelio, D. & Montresor, M. 2005. Life cycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae). *J. Phycol.* 41:542-56.

Amato, A. 2007. *The sexual cycle of the diatom Pseudo-nitzschia: from morphology through biology to gene.* Open University of London.

Amato, A., Kooistra, W. H. C. F., Levialdi Ghiron, J. H., Mann, D. G., Pröschold, T. & Montresor, M. 2007. Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 158:193-207.

Amato, A. & Montresor, M. 2008. Morphology, phylogeny, and sexual cycle of *Pseudo-nitzschia mannii* sp. nov. (Bacillariophyceae): a pseudo-cryptic species within the *P. pseudodelicatissima* complex. *Phycologia* 47:487–97.

Amin, S. A., Hmelo, L. R., van Tol, H. M., Durham, B. P., Carlson, L. T., Heal, K. R., Morales, R. L., Berthiaume, C. T., Parker, M. S., Djunaedi, B., Ingalls, A. E., Parsek, M. R., Moran, M. A. & Armbrust, E. V. 2015. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* 522:98-101.

Andersen, R. A. 2005. *Algal culturing techniques.* Elsevier, Academic Press, 578.

Armbrust, E. V., Chisholm, S. W. & Olson, R. J. 1990. Role of light and the cell cycle on the induction of spermatogenesis in a centric diatom. *J. Phycol.* 26:470-78.

Armbrust, E. V. 1999. Identification of a new gene family expressed during the onset of sexual reproduction in the centric diatom *Thalassiosira weissflogii*. *Appl. Environ. Microbiol.* 65:3121-28.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov,

V. V., Kröger, N., Lau, W. W. Y., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P. & Rokhsar, D. S. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79-86.

Armbrust, E. V. 2009. The life of diatoms in the world's oceans. *Nature* 459:185-92.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. & Eppig, J. T. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25-29.

Ashworth, J., Coesel, S., Lee, A., Armbrust, E. V., Orellana, M. V. & Baliga, N. S. 2013. Genome-wide diel growth state transitions in the diatom *Thalassiosira pseudonana. PNAS* 110:7518-23.

Assmy, P., Henjes, J., Smetacek, V. & Montresor, M. 2006. Auxospore formation in the silica-sinking oceanic diatom *Fragilariopsis kerguelensis* (Bacillariophyceae). *J. Phycol.* 42:1002-06.

Assmy, P., Hernández-Becerril, D. U. & Montresor, M. 2008. Morphological variability and life cycle traits of the type species of the diatom genus *Chaetoceros, C. dichaeta. J. Phycol.* 44:152-63.

Assmy, P. & Smetacek, V. 2009. Algal Blooms. *In*: Schaechter, M. [Ed.] *Encyclopedia of Microbiology.* Oxford, Elsevier, pp. 27-41.

Bachtrog, D., Kirkpatrick, M., Mank, J. E., McDaniel, S. F., Pires, J. C., Rice, W. & Valenzuela, N. 2011. Are all sex chromosomes created equal? *Trends Genet.* 27:350-57.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J. C. & The Tree of Sex, C. 2014. Sex determination: why so many ways of doing It? *PLoS Biol.* 12:e1001899.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* 28:304-05.

Baker, B. S. & Belote, J. M. 1983. Sex determination and dosage compensation in *Drosophila melanogaster. Annu. Rev. Genet.* 17:345-93.

Bartsch, I., Wiencke, C., Bischof, K., Buchholz, C. M., Buck, B. H., Eggert, A., Feuerpfeil, P., Hanelt, D., Jacobsen, S. & Karez, R. 2008. The genus *Laminaria sensu lato*: recent insights and developments. *Eur. J. Phycol.* 43:1-86.

Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., Maumus, F., Casotti, R., Mock, T., Caccamo, M., Montresor, M., Sanges, R. & Ferrante, M. I. under

revision. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom.

Bergero, R. & Charlesworth, D. 2009. The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* 24:94-102.

Bertrand, E. M., Allen, A. E., Dupont, C. L., Norden-Krichmar, T. M., Bai, J., Valas, R. E. & Saito, M. A. 2012. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *PNAS* 109:E1762–E71.

Bloomfield, G., Skelton, J., Ivens, A., Tanaka, Y. & Kay, R. R. 2010. Sex determination in the social amoeba *Dictyostelium discoideum*. *Science* 330:1533-36.

Boisson-Dernier, A., Roy, S., Kritsas, K., Grobei, M. A., Jaciubek, M., Schroeder, J. I. & Grossniklaus, U. 2009. Disruption of the pollen-expressed FERONIA homologs ANXUR1 and ANXUR2 triggers pollen tube discharge. *Development* 136:3279-88.

Botte, V., Ribera D'Alcalà, M. & Montresor, M. 2013. Hydrodynamic interactions at low Reynolds number: an overlooked mechanism favouring diatom encounters. *J. Plankton Res.* 35:914-18.

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C. J., Coesel, S., De Martino, A., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P. G., La Roche, J., Lindquist, E., Lommer, M., Martin–Jézéquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Marie-Pierre, Oudot–Le Secq, M.-P., Napoli, C., Obornik, M., Schnitzler Parker, M., Petit, J.-L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Van de Peer, Y. & Grigoriev, I. V. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239-44.

Bowler, C., Vardi, A. & Allen, A. E. 2010. Oceanographic and biogeochemical insights from diatom genomes. *An. Rev. Mar. Sci.* 2:333-65.

Brzezinski, M. A., Olson, R. J. & Chisholm, S. W. 1990. Silicon availability and cell-cycle progression in marine diatoms. *Mar. Ecol.-Prog. Ser.* 67:83-96.

Bull, J. J. 1983. *Evolution of sex determining mechanisms.* The Benjamin/Cummings Publishing Company, Inc., 316.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Casteleyn, G., Chepurnov, V. A., Leliaert, F., Mann, D. G., Bates, S. S., Lundholm, N., Rhodes, L., Sabbe, K. & Vyverman, W. 2008. *Pseudo-nitzschia pungens* (Bacillariophyceae): A cosmopolitan diatom species? *Harmful Algae* 7:241-57.

Casteleyn, G., Adams, N. G., Vanormelingen, P., Debeer, A.-E., Sabbe, K. & Vyverman, W. 2009. Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae): genetic and morphological evidence. *Protist* 160:343-54.

Cervantes, M. D., Hamilton, E. P., Xiong, J., Lawson, M. J., Yuan, D., Hadjithomas, M., Miao, W. & Orias, E. 2013. Selecting one of several mating types through gene segment joining and deletion in *Tetrahymena thermophila. PLoS Biol.* 11:e1001518.

Charlesworth, D. 2002. Plant sex determination and sex chromosomes. *Heredity* 88:94-101.

Charlesworth, D. 2013. Plant sex chromosome evolution. *J. Exp. Bot.* 64:405-20.

Chen, X., Mei, J., Wu, J., Jing, J., Ma, W., Zhang, J., Dan, C., Wang, W. & Gui, J.-F. 2015. A comprehensive transcriptome provides candidate genes for sex determination/differentiation and SSR/SNP markers in yellow catfish. *Mar. Biotechnol.* 17:190-98.

Chepurnov, V. A., Mann, D. G., Sabbe, K. & Vyverman, W. 2004. Experimental studies on sexual reproduction in diatoms. *Int. Rev. Cytol.* 237:91-154.

Chepurnov, V. A., Mann, D. G., Sabbe, K., Vannerum, K., Casteleyn, G., Verleyen, E., Peperzak, L. & Vyverman, W. 2005. Sexual reproduction, mating system, chloroplast dynamics and abrupt cell size reduction in *Pseudo-nitzschia pungens* from the North Sea (Bacillariophyta). *Eur. J. Phycol.* 40:379-95.

Chepurnov, V. A., Mann, D. G., von Dassow, P., Armbrust, E. V., Sabbe, K., Dasseville, R. & Vyverman, W. 2006. Oogamous reproduction, with two-step auxosporulation, in the centric diatom *Thalassiosira punctigera* (Bacillariophyta). *J. Phycol.* 42:845-58.

Chepurnov, V. A., Mann, D. G., von Dassow, P., Vanormelingen, P., Gillard, J., Inzé, D., Sabbe, K. & Vyverman, W. 2008. In search of new tractable diatoms for experimental biology. *BioEssays* 30:692–702.

Chi, J. Y., Parrow, M. W. & Dunthorn, M. 2014. Cryptic sex in *Symbiodinium* (Alveolata, Dinoflagellata) is supported by an inventory of meiotic genes. *J. Eukaryot. Microbiol.* 61:322-27.

Chisholm, S. W., Armbrust, E. V. & Olson, R. J. 1986. The individual cell in phytoplankton ecology: cell cycles and applications of flow cytometry. *In*: Platt, T. & Li, W. K. W. [Eds.] *Photosynthetic picoplankton.* Canadian Bullettin of Fisheries and Aquatic Sciences, Ottawa, pp. 343-69.

Churro, C. I., Carreira, C. C., Rodrigues, F. J., Craveiro, S. C., Calado, A. J., Casteleyn, G. & Lundholm, N. 2009. Diversity and abundance of potentially toxic *Pseudo-nitzschia* Peragallo in Aveiro coastal lagoon, Portugal and description of a new variety, *P. pungens* var. *aveirensis* var. nov. *Diatom Res.* 24:35-62.

Claesen, J., Clement, L., Shkedy, Z., Foulquié-Moreno, M. R. & Burzykowski, T. 2013. Simultaneous mapping of multiple gene loci with pooled segregants. *PLoS ONE* 8:e55133.

Coelho, S. M., Godfroy, O., Arun, A., Le Corguille, G., Peters, A. F. & Cock, J. M. 2011. Genetic regulation of life cycle transitions in the brown alga *Ectocarpus*. *Plant Sig. Behav.* 6:1858-60.

Crawford, R. M. 1995. The role of sex in the sedimentation of a marine diatom bloom. *Limnol. Oceanogr.* 40:200-04.

D'Alelio, D., Amato, A., Luedeking, A. & Montresor, M. 2009. Sexual and vegetative phases in the planktonic diatom *Pseudo-nitzschia multistriata*. *Harmful Algae* 8:225-32.

D'Alelio, D., Ribera d'Alcalà, M., Dubroca, L., Sarno, D., Zingone, A. & Montresor, M. 2010. The time for sex: a biennial life cycle in a marine planktonic diatom. *Limnol. Oceanogr.* 55:106-14.

Davidovich, N. A. & Bates, S. S. 1998. Sexual reproduction in the pennate diatoms *Pseudo-nitzschia multiseries* and *P. pseudodelicatissima* (Bacillariophyceae). *J. Phycol.* 34:126-37.

Davidovich, N. A., Ehrman, J. M. & Kaczmarska, I. 2006. The sexual structure of a natural population of the diatom *Nitzschia longissima* (Bréb.) Ralfs. *In*: Witkowski, A. [Ed.] *Proceedings of the 18th International Diatom Symposium.* Biopress Limited, Bristol, pp. 27-40.

Davidovich, N. A., Mouget, J.-L. & Gaudin, P. 2009. Heterothallism in the pennate diatom *Haslea ostrearia* (Bacillariophyta). *Eur. J. Phycol.* 44:251 - 61.

Davidovich, N. A., Gastineau, R., Gaudin, P., Davidovich, O. I. & Mouget, J. L. L. 2012. Sexual reproduction in the newly–described blue diatom, *Haslea karadagensis*. *Fottea* 12:219-29.

Davis, C. O., Harrison, P. J. & Dugdale, R. C. 1973. Continuous culture of marine diatoms under silicate limitation. I. Synchronized life cycle of *Skeletonema costatum*. *J. Phycol.* 9:175-80.

De Riso, V., Raniello, R., Maumus, F., Rogato, A., Bowler, C. & Falciatore, A. 2009. Gene silencing in the marine diatom *Phaeodactylum tricornutum*. *Nucleic Acids Res.* 37.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T. O., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P. & Karsenti, E. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605-11.

Derenbach, J. B. & Pesando, D. 1986. Investigations into a small fraction of volatile hydrocarbons: III. 2 Diatom cultures produce ectocarpene, a pheromone of brown-algae. *Mar. Chem.* 19:337-41.

Derveaux, S., Vandesompele, J. & Hellemans, J. 2010. How to do successful gene expression analysis using real-time PCR. *Methods* 50:227-30.

Di Dato, V., Musacchia, F., Petrosino, G., Patil, S., Montresor, M., Sanges, R. & Ferrante, M. I. 2015. Transcriptome sequencing of three *Pseudo-nitzschia* species reveals comparable gene sets and the presence of Nitric Oxide Synthase genes in diatoms. *Sci. Rep.* 5:12329.

Drebes, G. 1964. Über den lebenszyklus der marinen planktondiatomee *Stephanopyxis turris* (Centrales) und seine steuerung im experiment. *Helgol. Wiss. Meeresunters.* 10:153-54.

Drebes, G. 1966. On the life history of the marine plankton diatom *Stephanopyxis palmeriana*. *Helgol. Wiss. Meeresunters.* 13:104-14.

Drebes, G. 1972. The life history of the centric diatom *Bacteriastrum hyalinum* Lauder. *Nova Hedwigia Beih.* 39:95-110.

Drebes, G. 1977a. Cell structure, cell division and sexual reproduction of *Attheya decora* West (Bacillariophyceae, Bidduphineae). *Nova Hedwigia Beih.* 54:167-78.

Drebes, G. 1977b. Sexuality. *In*: Werner, D. [Ed.] *The biology of diatoms.* Blackwell Scientific Publications, Oxford, pp. 250-83.

Drebes, G. 1979. Oogame auxosporenbildung bei *Thalassiosira eccentrica. Jahresbericht Biologische Anstalt Helgoland* 5.

Dyhrman, S. T., Jenkins, B. D., Rynearson, T. A., Saito, M. A., Mercier, M. L., Alexander, H., Whitney, L. P., Drzewianowski, A., Bulygin, V. V., Bertrand, E. M., Wu, Z., Benitez-Nelson, C. & Heithoff, A. 2012. The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. *PLoS ONE* 7:e33768.

Ekblom, R. & Galindo, J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107:1-15.

Ellegren, H. & Parsch, J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Gen.* 8:689-98.

Escobar-Restrepo, J.-M., Huck, N., Kessler, S., Gagliardini, V., Gheyselinck, J., Yang, W.-C. & Grossniklaus, U. 2007. The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. *Science* 317:656-60.

Ferris, P., Olson, B. J. S. C., De Hoff, P. L., Douglass, S., Casero, D., Prochnik, S., Geng, S., Rai, R., Grimwood, J., Schmutz, J., Nishii, I., Hamaji, T., Nozaki, H., Pellegrini, M. & Umen, J. G. 2010. Evolution of an expanded sex-determining locus in *Volvox. Science* 328:351-54.

Ferris, P. J. & Goodenough, U. W. 1994. The mating-type locus of *Chlamydomonas reinhardtii* contains highly rearranged DNA sequences. *Cell* 76:1135-45.

Field, H. I., Coulson, R. M. & Field, M. C. 2013. An automated graphics tool for comparative genomics: the Coulson plot generator. *BMC Bioinformatics* 14:1-8.

Fistarol, G. O., Legrand, C., Rengefors, K. & Granéli, E. 2004. Temporary cyst formation in phytoplankton: a response to allelopathic competitors? *Environ. Microbiol.* 6:791-98.

Fleige, S. & Pfaffl, M. W. 2006. RNA integrity and the effect on the real-time qRT-PCR performance. *Molecular Aspects of Medicine* 27:126-39.

Foster, R. & Kreitzman, L. 2004. *The rhythms of life: the biological clocks that control the daily lives of every living thing.* Yale University Press, 276.

Fraser, J. A., Diezmann, S., Subaran, R. L., Allen, A., Lengeler, K. B., Dietrich, F. S. & Heitman, J. 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biol.* 2:e384.

Fraser, J. A. & Heitman, J. 2005. Chromosomal sex-determining regions in animals, plants and fungi. *Curr. Opin. Genet. Dev.* 15:645-51.

Fraser, J. A., Stajich, J. E., Tarcha, E. J., Cole, G. T., Inglis, D. O., Sil, A. & Heitman, J. 2007. Evolution of the mating type locus: insights gained from the dimorphic primary fungal pathogens *Histoplasma capsulatum*, *Coccidioides immitis*, and *Coccidioides posadasii*. *Eukaryot. Cell* 6:622-29.

French III, F. W. & Hargraves, P. E. 1985. Spore formation in the life cycles of the diatoms *Chaetoceros diadema* and *Leptocylindrus danicus*. *J. Phycol.* 21:477-83.

Frenkel, J., Vyverman, W. & Pohnert, G. 2014. Pheromone signaling during sexual reproduction in algae. *The Plant Journal* 79:632-44.

Fryxell, G. A., Garza, S. A. & Roelke, D. L. 1991. Auxospore formation in an Antarctic clone of *Nitzschia subcurvata* Hasle. *Diatom Res.* 6:235-45.

Fuchs, N., Scalco, E., Kooistra, W. C. H. F., Assmy, P. & Montresor, M. 2013. Genetic characterization and life cycle of the diatom *Fragilariopsis kerguelensis*. *Eur. J. Phycol.* 48:411-26.

Furnas, M. J. 1985. Diel synchronization of sperm formation in the diatom *Chaetoceros curvisetus* Cleve. *J. Phycol.* 21:667-71.

Gallagher, J. C. 1983. Cell enlargement in *Skeletonema costatum* (Bacillariophyceae). *J. Phycol.* 19:539-42.

Gastineau, R., Leignel, V., Jacquette, B., Hardivillier, Y., Wulff, A., Gaudin, P., Bendahmane, D., Davidovich, N. A., Kaczmarska, I. & Mouget, J.-L. 2013. Inheritance of mitochondrial DNA in the pennate diatom *Haslea ostrearia* (Naviculaceae) during auxosporulation suggests a uniparental transmission. *Protist* 164:340-51.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. & Nekrutenko, A. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* 15:1451-55.

Gillard, J., Devos, V., Huysman, M. J. J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V. A., Inze, D., Vuylsteke, M. & Vyverman, W. 2008. Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiol.* 148:1394-411.

Gillard, J., Frenkel, J., Devos, V., Sabbe, K., Paul, C., Rempt, M., Inze, D., Pohnert, G., Vuylsteke, M. & Vyverman, W. 2013. Metabolomics enables the structure elucidation of a diatom sex pheromone. *Angew. Chem. Int. Ed.* 52:854-57.

Gissendanner, C. & Kelley, T. 2013. The *C. elegans* gene pan-1 encodes novel transmembrane and cytoplasmic leucine-rich repeat proteins and promotes molting and the larva to adult transition. *BMC Dev. Biol.* 13:21.

Godhe, A., Kremp, A. & Montresor, M. 2014. Genetic and microscopic evidence for sexual reproduction in the centric diatom *Skeletonema marinoi*. *Protist* 165:401-16.

Goodenough, U., Lin, H. & Lee, J.-H. 2007. Sex determination in *Chlamydomonas*. *Semin. Cell Dev. Biol.* 18:350-61.

Goodenough, U. W., Armbrust, E. V., Campbell, A. M. & Ferris, P. J. 1995. Molecular genetics of sexuality in *Chlamydomonas*. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 44:21-44.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R. & Zeng, Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644-52.

Grimsley, N., Pequin, B., Bachy, C., Moreau, H. & Piganeau, G. 2010. Cryptic sex in the smallest eukaryotic marine green alga. *Mol. Biol. Evol.* 27:47-54.

Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. & Mock, T. 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *The Plant Journal* 81:519-28.

Guillard, R. R. L. 1975. Culture of phytoplankton for feeding marine invertebrates. *In*: Smith, W. L. & Chanley, M. H. [Eds.] *Culture of Marine Invertebrate Animals.* Plenum Press, New York, pp. 29-60.

Haag, E. S. & Doty, A. V. 2005. Sex determination across evolution: connecting the dots. *PLoS Biol.* 3:e21.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B. & Lieber, M. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8:1494-512.

Hamaji, T., Ferris, P. J., Nishii, I., Nishimura, Y. & Nozaki, H. 2013. Distribution of the sex-determining gene *MID* and molecular correspondence of mating types within the isogamous genus *Gonium* (Volvocales, Chlorophyta). *PLoS ONE* 8:e64385.

Handley, C. J., Mok, M. T., Ilic, M. Z., Adcocks, C., Buttle, D. J. & Robinson, H. C. 2001. Cathepsin D cleaves aggrecan at unique sites within the interglobular domain and chondroitin sulfate attachment regions that are also cleaved when cartilage is maintained at acid pH. *Matrix Biol.* 20:543-53.

Hardege, J. D., Bartels-Hardege, H., Müller, C. T. & Beckmann, M. 2004. Peptide pheromones in female *Nereis succinea*. *Peptides* 25:1517-22.

Hargraves, P. E. 1972. Studies on marine plankton diatoms. I. *Chaetoceros diadema* (Ehr) Gran, life cycle, structural morphology and regional distribution. *Phycologia* 11:247-57.

Hargraves, P. E. 1976. Studies on marine plankton diatoms. III. Structure and classification of *Gossleriella tropica*. *J. Phycol.* 12:285-91.

Hay, M. E. 2009. Marine chemical ecology: chemical signals and cues structure marine populations, communities, and ecosystems. *An. Rev. Mar. Sci.* 1:193-212.

Hegde, S., Narale, D. D. & Anil, A. C. 2011. Sexual reproduction in *Odontella regia* (Schultze) Simonsen 1974 (Bacillariophyta). *Curr. Sci.* 101:222-25.

Hilbe, J. M. 2007. STATISTICA 7. *The American Statistician* 61:91-94.

Hiltz, M., Bates, S. S. & Kaczmarska, I. 2000. Effect of light:dark cycles and cell apical length on the sexual reproduction of the pennate diatom *Pseudo-nitzschia multiseries* (Bacillariophyceae) in culture. *Phycologia* 39:59-66.

Hirano, N., Marukawa, Y., Abe, J., Hashiba, S., Ichikawa, M., Tanabe, Y., Ito, M., Nishii, I., Tsuchikane, Y. & Sekimoto, H. 2015. A receptor-like kinase, related with cell wall sensor of higher plants, is required for sexual reproduction in the unicellular charophycean alga, *Closterium peracerosum-strigosum-littorale* complex. *Plant Cell Physiol.* 56:1456-62.

Holmes, R. W. 1966. Short-term temperature and light conditions associated with auxospore formation in the marine centric diatom *Coscinodiscus concinnus* W. Smith. *Nature* 209:217-18.

Holtermann, K. E., Bates, S. S., Trainer, V. L., Odell, A. & Armbrust, E. V. 2010. Mass sexual reproduction in the toxigenic diatoms *Pseudo-nitzschia australis* and *P. pungens* (Bacillariophyceae) on the Washington coast. *J. Phycol.* 46:41-52.

Honda, D., Shono, T., Kimura, K., Fujita, S., Iseki, M., Makino, Y. & Murakami, A. 2007. Homologs of the sexually induced gene 1 (sig1) product constitute the stramenopile mastigonemes. *Protist* 158:77-88.

Hoppenrath, M., Elbrächter, M. & Drebes, G. 2009. *Marine Phytoplankton. Selected microphytoplankton species from the North Sea around Helgoland and Sylt.* Kleine Senckenberg-Reihe, 264.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. & Nakai, K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35:W585-W87.

Hsu, A.-L., Murphy, C. T. & Kenyon, C. 2003. Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science* 300:1142-45.

Hurst, L. D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486-87.

Huysman, M. J., Vyverman, W. & De Veylder, L. 2013. Molecular regulation of the diatom cell cycle. *J. Exp. Bot.* 65:2573-84.

Huysman, M. J. J., Martens, C., Vandepoele, K., Gillard, J., Rayko, E., Heijde, M., Bowler, C., Inze, D., Van de Peer, Y., De Veylder, L. & Vyverman, W. 2010. Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome Biol.* 11:r17.

Hwang, Y.-s., Jung, G. & Jin, E. 2008. Transcriptome analysis of acclimatory responses to thermal stress in Antarctic algae. *Biochem. Biophys. Res. Commun.* 367:635-41.

Idei, M. & Chihara, M. 1992. Successive observations on the fertilization of a centric diatom *Melosira moniliformis* var. *octagona*. *Bot. Mag. Tokyo* 105:649–58.

Idnurm, A. 2011. Sex determination in the first-described sexual fungus. *Eukaryot. Cell* 10:1485-91.

Ingleby, F. C., Flis, I. & Morrow, E. H. 2014. Sex-biased gene expression and sexual conflict throughout development. *Cold Spring Harbor Perspect. Biol.* 7:a017632.

Irish, E. E. & Nelson, T. 1989. Sex determination in monoecious and dioecious plants. *The Plant Cell* 1:737.

Jensen, K. G., Moestrup, Ø. & Schmid, A.-M. M. 2003. Ultrastructure of the male gametes from two centric diatoms, *Chaetoceros laciniosus* and *Coscinodiscus wailesii* (Bacillariophyceae). *Phycologia* 42:98-105.

Kaczmarska, I., Bates, S. S., Ehrman, J. M. & Léger, C. 2000. Fine structure of the gamete, auxospore and initial cell in the pennate diatom *Pseudo-nitzschia multiseries* (Bacillariophyta). *Nova Hedwigia* 71:337-57.

Kaczmarska, I., Davidovich, N. A. & Ehrman, J. M. 2007. Sex cells and reproduction in the diatom *Nitzschia longissima* (Bacillariophyta): discovery of siliceous scales in gamete cell walls and novel elements of the perizonium. *Phycologia* 46:726-37.

Kaczmarska, I., Reid, C., Martin, J. L. & Moniz, M. B. J. 2008. Morphological, biological, and molecular characteristics of the diatom *Pseudo-nitzschia delicatissima* from the Canadian Maritimes. *Botany* 86:763-72.

Kaczmarska, I., Poulíčková, A., Sato, S., Edlund, M. B., Idei, M., Watanabe, T. & Mann, D. G. 2013. Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. *Diatom Res.*:1-32.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K. & Bell, C. J. 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* 12:e1001889.

Klar, A. J. S. 2010. The yeast mating-type switching mechanism: a memoir. *Genetics* 186:443-49.

Kobiyama, A., Ikeda, Y., Koike, K. & Ogata, T. 2007. Isolation of a differentially expressed gene in separate mating types of the dinoflagellate *Alexandrium tamarense*. *Eur. J. Phycol.* 42:183-90.

Koester, J. A., Brawley, S. H., Karp-Boss, L. & Mann, D. G. 2007. Sexual reproduction in the marine centric diatom *Ditylum brightwellii* (Bacillariophyta). *Eur. J. Phycol.* 42:351-66.

Kooistra, W. H. C. F., De Stefano, M., Medlin, L. K. & Mann, D. G. 2003. The phylogeny of the diatoms. *In*: Müller, W. E. G. [Ed.] *Silicon Biomineralization.* Progress in molecular and subcellular biology, pp. 59-97.

Kooistra, W. H. C. F., Gersonde, R., Medlin, L. K. & Mann, D. G. 2007. The origin and evolution of the diatoms: their adaptation to a planktonic existence. *In*: Falkowski, P. G. & Knoll, A. H. [Eds.] *Evolution of Primary Producers in the Sea* Elsevier Academic Press, Burlington, pp. 207-50.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B. & Strömbom, L. 2006. The real-time polymerase chain reaction. *Molecular aspects of medicine* 27:95-125.

Kurihara, M. & Takahashi, K. 2002. Long-term size variation and life cycle patterns of a predominant diatom *Neodenticula seminae* in the Subarctic Pacific and Bering Sea. *Bull. Plankton Soc. Jpn.* 49:77-87.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Lauritano, C., Orefice, I., Procaccini, G., Romano, G. & Ianora, A. 2015. Key genes as stress indicators in the ubiquitous diatom *Skeletonema marinoi*. *BMC Genomics* 16:411.

Lechuga-Devéze, C. H. & Hernández-Becerril, D. U. 1988. Life cycle of the diatom *Chaetoceros protuberans* Lauder (1864) (Bacillariophyceae). *Invest. Pesq.* 52:77-83.

Letunic, I., Doerks, T. & Bork, P. 2015. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* 43:D257-D60.

Levialdi Ghiron, J. H., Amato, A., Montresor, M. & Kooistra, W. C. H. F. 2008. Plastid inheritance in the planktonic raphid pennate diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae). *Protist* 159:91-98.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing Subgroup 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078-79.

Lipinska, A., D'hondt, S., Van Damme, E. & De Clerck, O. 2013. Uncovering the genetic basis for early isogamete differentiation: a case study of Ectocarpus siliculosus. *BMC Genomics* 14:909.

Lipinska, A., Cormier, A., Luthringer, R., Peters, A. F., Corre, E., Gachon, C. M. M., Cock, J. M. & Coelho, S. M. 2015a. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *Ectocarpus*. *Mol. Biol. Evol.* 32:1581-97.

Lipinska, A. P., Ahmed, S., Peters, A. F., Faugeron, S., Cock, J. M. & Coelho, S. M. 2015b. Development of PCR-based markers to determine the sex of kelps. *PLoS ONE* 10:e0140535.

Lohse, M., Bolger, A., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M. & Usadel, B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.*:W622-W27.

Lommer, M., Specht, M., Roy, A.-S., Kraemer, L., Andreson, R., Gutowska, M., Wolf, J., Bergner, S., Schilhabel, M., Klostermeier, U., Beiko, R., Rosenstiel, P., Hippler, M. & LaRoche, J. 2012. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* 13:R66.

Lundholm, N., Hasle, G. R., Fryxell, G. A. & Hargraves, P. E. 2002. Morphology, phylogeny and taxonomy of species within the *Pseudo-nitzschia americana* complex (Bacillariophyceae) with descriptions of the two new species *P. brasiliana* and *P. linea*. *Phycologia* 41:480-97.

Lundholm, N., Moestrup, Ø., Hasle, G. R. & Hoef-Emden, K. 2003. A study of the *Pseudo-nitzschia pseudodelicatissima/cuspidata* complex (Bacillariophyceae): what is *P. pseudodelicatissima*? *J. Phycol.* 39:797-813.

Lundholm, N., Moestrup, Ø., Kotaki, Y., Hoef-Emden, K., Scholin, C. & Miller, P. 2006. Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima* complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. *J. Phycol.* 42:464-81.

Lundholm, N., Bates, S. S., Baugh, K. A., Bill, B. D., Connell, L. B., Léger, C. & Trainer, V. L. 2012. Cryptic and pseudo-cryptic diversity in diatoms—with descriptions of

Pseudo-nitzschia hasleana sp. nov. and *P. fryxelliana* sp. nov. *J. Phycol.* 48:436-54.

Lyczkowski, E. R. & Karp-Boss, L. 2014. Allelopathic effects of *Alexandrium fundyense* (Dinophyceae) on *Thalassiosira* cf. *gravida* (Bacillariophyceae): a matter of size. *J. Phycol.* 50:376-87.

MacIntyre, H. L. & Cullen, J. J. 2005. Using cultures to investigate the physiological ecology of microalgae. *In*: Andersen, R. A. [Ed.] *Algal Culturing Techniques.* Elsevier Academic Press, Burlington, MA, USA, pp. 287-326.

Magwene, P. M., Willis, J. H. & Kelly, J. K. 2011. The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comp. Biol.* 7:e1002255.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A. & Bowler, C. 2016. Insights into global diatom distribution and diversity in the world's ocean. *PNAS* 113:E1516-E25.

Mann, D. G., Chepurnov, V. A. & Droop, S. J. M. 1999. Sexuality, incompatibility, size variation, and preferencial polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). *J. Phycol.* 35:152-70.

Mann, D. G. & Vanormelingen, P. 2013. An inordinate fondness? the number, distributions,and origins of diatom species. *J. Euk. Microbiol.* 60:414–20.

Manton, I. & von Stoch, H. A. 1966. Observations on the fine structure of the male gamete of the marine centric diatom *Lithodesmium undulatum. J. R. Microsc. Soc.* 85:119-34.

Manton, I., Kowallik, K. & von Stosch, H. A. 1969. Observations on the fine structure and development of the spindle at mitosis and meiosis in a marine centric diatom (*Lithodesmium undulatum*). II. The early meiotic stages in male gametogenesis. *J. Cell Sci.* 5:271–98.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C. & Gonzales, N. R. 2011. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 39:D225-D29.

Marin, R. & Tanguay, R. 1996. Stage-specific localization of the small heat shock protein Hsp27 during oogenesis in *Drosophila melanogaster. Chromosoma* 105:142-49.

Martins, M. J. F., Mota, C. & Pearson, G. 2013. Sex-biased gene expression in the brown alga *Fucus vesiculosus. BMC Genomics* 14:294.

Matsunaga, S. & Kawano, S. 2001. Sex determination by sex chromosomes in dioecious plants. *Plant Biol.* 3:481-88.

McQuoid, M. R. & Hobson, L. A. 1996. Diatom resting stages. *J. Phycol.* 32:889-902.

Merlini, L., Dudin, O. & Martin, S. G. 2013. Mate and fuse: how yeast cells do it. *Open Biol.* 3:130008.

Metin, B., Findley, K. & Heitman, J. 2010. The mating type locus (*MAT*) and sexual reproduction of *Cryptococcus heveanensis*: insights into the evolution of sex and sex-determining chromosomal regions in fungi. *PLoS Genet.* 6:e1000961.

Migita, S. 1967a. Sexual reproduction of centric diatom *Skeletonema costatum*. *Bull. Jap. Soc. Sci. Fish.* 33:392-98.

Migita, S. 1967b. Sexual reproduction of *Melosira moniliformis* Agardh. *Bull. Fac. Fish. Nagasaki Univ.* 23:123–33.

Mills, K. E. & Kaczmarska, I. 2006. Autogamic reproductive behavior and sex cell structure in *Thalassiosira angulata* (Bacillariophyta). *Bot. Mar.* 49:417-30.

Ming, R., Bendahmane, A. & Renner, S. S. 2011. Sex chromosomes in land plants. *Annu. Rev. Plant Biol.* 62:485-514.

Miyazaki, S., Murata, T., Sakurai-Ozato, N., Kubo, M., Demura, T., Fukuda, H. & Hasebe, M. 2009. ANXUR1 and 2, sister genes to FERONIA/SIRENE, are male factors for coordinated fertilization. *Curr. Biol.* 19:1327-31.

Moeys, S., Frenkel, J., Lembke, C., Gillard, J. T., Devos, V., Van den Berge, K., Bouillon, B., Huysman, M. J., De Decker, S. & Scharf, J. 2016. A sex-inducing pheromone triggers cell cycle arrest and mate attraction in the diatom *Seminavis robusta*. *Sci. Rep.* 6:19252

Montresor, M., Vitale, L., D'Alelio, D. & Ferrante, M. I. 2016. Sex in marine planktonic diatoms: insights and challenges. *Perspec. Phycol.* 3:61-75.

Montsant, A., Allen, A. E., Coesel, S., Martino, A. D., Falciatore, A., Mangogna, M., Siaut, M., Heijde, M., Jabbari, K., Maheswari, U., Rayko, E., Vardi, A., Apt, K. E., Berges, J. A., Chiovitti, A., Davis, A. K., Thamatrakoln, K., Hadi, M. Z., Lane, T. W., Lippmeier, J. C., Martinez, D., Parker, M. S., Pazour, G. J., Saito, M. A., Rokhsar, D. S., Armbrust, E. V. & Bowler, C. 2007. Identification and comparative genomic analysis of signaling and regulatory components in the diatom *Thalassiosira pseudonana*. *J. Phycol.* 43:585-604.

Morozova, O., Hirst, M. & Marra, M. A. 2009. Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genom. Hum. Genet.* 10:135-51.

Mouget, J. L., Gastineau, R., Davidovich, O., Gaudin, P. & Davidovich, N. A. 2009. Light is a key factor in triggering sexual reproduction in the pennate diatom *Haslea ostrearia*. *FEMS Microbiol. Ecol.* 69:194-201.

Muramoto, T. & Urushihara, H. 2006. Small GTPase RacF2 affects sexual cell fusion and asexual development in *Dictyostelium discoideum* through the regulation of cell adhesion. *Dev. Growth Differ.* 48:199-208.

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M. & Sanges, R. 2015. Annocript: a flexible pipeline for the annotation of transcriptomes also able to identify putative long noncoding RNAs. *Bioinformatics* 31:2199-201.

Nagai, S. & Manabe, T. 1994. Auxospore formation of a giant diatom, *Coscinodiscuc wailesii* (Bacillariophyceae), in culture. *Bull. Plankt. Soc. Japan* 40:151-67.

Nagai, S., Imai, I., Yamauchi, K. & Manabe, T. 1996. Induction of sexuality in the diatom *Coscinodiscus wailesii* Gran by a marine bacterium *Alcaligenes* sp. in culture. *In*: Mayama, I. & Koizumi, I. [Eds.] *14th Diatom Symposium.* Koeltz Scientific Books, Koeningstein, pp. 198-212.

Nanjappa, D., Kooistra, W. H. C. F. & Zingone, A. 2013. A reappraisal of the genus *Leptocylindrus* (Bacillariophyta), with the addition of three species and the erection of *Tenuicylindrus* gen. nov. . *J. Phycol.* 49:917–36.

Nekrutenko, A., Makova, K. D. & Li, W.-H. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12:198-202.

Neuer, A., Spandorfer, S., Giraldo, P., Dieterle, S., Rosenwaks, Z. & Witkin, S. 2000. The role of heat shock proteins in reproduction. *Hum. Reprod. Update* 6:149-59.

Nishikawa, T., Hori, Y., Harada, K. & Imai, I. 2013. Annual regularity of reduction and restoration of cell size in the harmful diatom *Eucampia zodiacus*, and its application to the occurrence prediction of nori bleaching. *Plankton Benthos Res.* 8:166-80.

Olson, R. J., Vaulot, D. & Chisholm, S. W. 1986. Effects of environmental stresses on the cell cycle of two marine phytoplankton species. *Plant Physiol.* 80:918-25.

Oppliger, L. V., Correa, J. A., Faugeron, S., Beltran, J., Tellier, F., Valero, M. & Destombe, C. 2011. Sex ratio variation in the *Lessonia nigrescens* complex (Laminariales, Phaeophyceae): effect of latitude, temperature and marginality. *J. Phycol.* 47:5-12.

Orsini, L., Sarno, D., Procaccini, G., Poletti, R., Dahlmann, J. & Montresor, M. 2002. Toxic *Pseudo-nitzschia multistriata* (Bacillariophyceae) from the Gulf of Naples: morphology, toxin analysis and phylogenetic relationships with other *Pseudo-nitzschia* species. *Eur. J. Phycol.* 37:247-57.

Palaiokostas, C., Bekaert, M., Khan, M. G. Q., Taggart, J. B., Gharbi, K., McAndrew, B. J. & Penman, D. J. 2013. Mapping and validation of the major sex-determining region in *Nile tilapia* using RAD sequencing. *PLoS ONE* 8: e68389.

Pan, J. & Snell, W. J. 2000. Signal transduction during fertilization in the unicellular green alga, *Chlamydomonas*. *Curr. Opin. Microbiol.* 3:596-602.

Park, S., Jung, G., Hwang, Y.-S. & Jin, E. 2010. Dynamic response of the transcriptome of a psychrophilic diatom, *Chaetoceros neogracile*, to high irradiance. *Planta* 231:349-60.

Parsch, J. & Ellegren, H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat. Rev. Gen.* 14:83-87.

Patil, S., Moeys, S., von Dassow, P., Huysman, M. J., Mapleson, D., De Veylder, L., Sanges, R., Vyverman, W., Montresor, M. & Ferrante, M. I. 2015. Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species *Pseudo-nitzschia multistriata* and *Seminavis robusta*. *BMC Genomics* 16:930.

Patil, S. M. 2014. *Genomics enable exploration in the marine planktonic diatom Genus Pseudo-nitzschia.* PhD, The Open University.

Paul, C., Barofsky, A., Vidoudez, C. & Pohnert, G. 2009. Diatom exudates influence metabolism and cell growth of co-cultured diatom species. *Mar. Ecol.-Prog. Ser.* 389:61-70.

Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8:785-86.

Pfaffl, M. W., Horgan, G. W. & Dempfle, L. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res.* 30:e36-e36.

Pohnert, G. & Boland, W. 2002. The oxylipin chemistry of attraction and defense in brown algae and diatoms. *Nat. Prod. Rep.* 19:108-22.

Quijano-Scheggia, S., Garcés, E., Andree, K., Fortuño, J. M. & Camp, J. 2009a. Homothallic auxosporulation in *Pseudo-nitzschia brasiliana* (Bacillariophyta). *J. Phycol.* 45:100-07.

Quijano-Scheggia, S. I., Garcés, E., Lundholm, N., Moestrup, O., Andree, K. & Camp, J. 2009b. Morphology, physiology, molecular phylogeny and sexual compatibility of the cryptic *Pseudo-nitzschia delicatissima* complex (Bacillariophyta), including the description of *P. arenysensis* sp. nov. *Phycologia* 48:492-509.

Rai, U. & Roy, B. 2008. Sex determination and differentiation in vertebrates. *In*: Misro, M. M. [Ed.] *Reproduction Biology.* National Institute of Health & Family Welfare (NIHFW), New Delhi.

Rayko, E., Maumus, F., Maheswari, U., Jabbari, K. & Bowler, C. 2010. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* 188:52-66.

Raymond, C. S., Shamu, C. E., Shen, M. M., Seifert, K. J., Hirsch, B., Hodgkin, J. & Zarkower, D. 1998. Evidence for evolutionary conservation of sex-determining genes. *Nature* 391:691-95.

Rittschof, D. & Cohen, J. H. 2004. Crustacean peptide and peptide-like pheromones and kairomones. *Peptides* 25:1503-16.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24-26.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-40.

Round, F. E. 1972. The problem of reduction of cell size during diatom cell division. *Nova Hedwigia* 23:291-303.

Round, F. E., Crawford, R. M. & Mann, D. G. 1990. *The diatoms. Biology and morphology of the genera.* Cambridge University Press, Cambridge, 747.

Russo, M. T., Annunziata, R., Sanges, R., Ferrante, M. I. & Falciatore, A. 2015. The upstream regulatory sequence of the light harvesting complex Lhcf2 gene of the marine diatom *Phaeodactylum tricornutum* enhances transcription in an orientation- and distance-independent fashion. *Marine Genomics* 24, Part 1:69-79.

Sabatino, V., Russo, M. T., Patil, S., d'Ippolito, G., Fontana, A. & Ferrante, M. I. 2015. Establishment of genetic transformation in the sexually reproducing diatoms *Pseudo-nitzschia multistriata* and *Pseudo-nitzschia arenysensis* and inheritance of the transgene. *Mar. Biotechnol.* 17:452-62.

Sarno, D., Zingone, A. & Montresor, M. 2010. A massive and simultaneous sex event of two *Pseudo-nitzschia* species. *Deep-Sea Res. Pt II* 57:248-55.

Sato, S., Beakes, G., Idei, M., Nagumo, T. & Mann, D. G. 2011. Novel sex cells and evidence for sex pheromones in diatoms. *PLoS ONE* 6:e26923.

Scalco, E. 2013. *Factors regulating transitions among life cycle phases in the marine pennate diatom Pseudo-nitzschia multistriata.* Open University, London, 274 pp.

Scalco, E., Stec, K., Iudicone, D., Ferrante, M. I. & Montresor, M. 2014. The dynamics of sexual phase in the marine diatom *Pseudo-nitzschia multistriata* (Bacillariophyceae). *J. Phycol.* 50:817-28.

Scalco, E., Amato, A., Ferrante, M. I. & Montresor, M. 2015. The sexual phase of the diatom *Pseudo-nitzschia multistriata*: cytological and time-lapse cinematography characterization. *Protoplasma*:1-11.

Schmid, A. M. M. 1995. Sexual reproduction in *Coscinodiscus granii* Gough in culture: a preliminary report. *In*: Marino, D. & Montresor, M. [Eds.] *Proceedings of the 13th international diatom symposium 1994.* Biopress, Bristol, pp. 139-59.

Schmittgen, T. D. & Livak, K. J. 2008. Analyzing real-time PCR data by the comparative CT method. *Nat. Protocols* 3:1101-08.

Schulze, B., Buhmann, M. T., Río Bártulos, C. & Kroth, P. G. 2015. Comprehensive computational analysis of leucine-rich repeat (LRR) proteins encoded in the genome of the diatom *Phaeodactylum tricornutum*. *Marine Genomics* 21:43-51.

Sekimoto, H., Tanabe, Y., Takizawa, M., Ito, N., Fukumoto, R. & Ito, M. 2003. Expressed sequence tags from the *Closterium peracerosum-strigosum-littorale* complex, a unicellular charophycean alga, in the sexual reproduction process. *DNA Res.* 10:147-53.

Sekimoto, H., Tanabe, Y., Tsuchikane, Y., Shirosaki, H., Fukuda, H., Demura, T. & Ito, M. 2006. Gene expression profiling using cDNA microarray analysis of the sexual reproduction stage of the unicellular charophycean alga *Closterium peracerosum-strigosum-littorale* complex. *Plant Physiol.* 141:271-79.

Sekimoto, H., Abe, J. & Tsuchikane, Y. 2012. New insights into the regulation of sexual reproduction in *Closterium*. *In*: Jeon, K. W. [Ed.] *International Review of Cell and Molecular Biology.* pp. 309-38.

Sekimoto, H., Tsuchikane, Y. & Abe, J. 2014. Sexual reproduction of a unicellular Charophycean alga, *Closterium peracerosum-strogosum-littorale* complex. *In*: Sawada, H., Inoue, N. & Iwano, M. [Eds.] *Sexual reproduction in animals and plants.* Springer Open, Tokyo, pp. 345-59.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. & Bowler, C. 2007. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* 406:23-35.

Smetacek, V. 1999. Revolution in the ocean. *Nature* 401:647.

Sournia, A., Chretiennot-Dinet, M. J. & Ricard, M. 1991. Marine phytoplankton:how many species in the world ocean? *J. Plankton Res.* 13:1093-99.

Star, B., Tørresen, O. K., Nederbragt, A. J., Jakobsen, K. S., Pampoulie, C. & Jentoft, S. 2016. Genomic characterization of the Atlantic cod sex-locus. *Sci. Rep.* 6:31235.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23:1282-88.

Takano, H. 1993. Marine diatom *Nitzschia multistriata* sp. nov. common at inlets of southern Japan. *Diatom* 8:39-41.

Takano, H. 1995. *Pseudo-nitzschia multistriata* (Takano) Takano, a new combination for the pennate diatom *Nitzschia multistriata* Takano. *Diatom* 10:73-74.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30:2725-29.

Tanaka, T., Maeda, Y., Veluchamy, A., Tanaka, M., Abida, H., Maréchal, E., Bowler, C., Muto, M., Sunaga, Y. & Tanaka, M. 2015. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *The Plant Cell* 27:162-76.

Tanurdzic, M. & Banks, J. A. 2004. Sex-determining mechanisms in land plants. *The Plant Cell* 16:S61-S71.

Tesson, S. V. N., Legrand, C., van Oosterhout, C., Montresor, M., Kooistra, W. H. C. F. & Procaccini, G. 2013. Mendelian inheritance pattern and high mutation rates of microsatellite alleles in the diatom *Pseudo-nitzschia multistriata. Protist* 164:89-100.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-80.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14:178-92.

Tillmann, U. & John, U. 2002. Toxic effects of *Alexandrium* spp. on heterotrophic dinoflagellates: an allelochemical defence mechanism independent of PSP-toxin content. *Mar. Ecol.-Prog. Ser.* 230:47-58.

Treguer, P., Nelson, D. M., Bennekom, A. J. V., Demaster, D. J., Leynaert, A. & Queguiner, B. 1995. The silica balance in the world ocean: a reestimate. *Science* 268:375-79.

Tsuchikane, Y., Kokubun, Y. & Sekimoto, H. 2010. Characterization and molecular cloning of conjugation-regulating sex pheromones in homothallic *Closterium*. *Plant Cell Physiol.* 51:1515-23.

Uller, T., Pen, I., Wapstra, E., Beukeboom, L. W. & Komdeur, J. 2007. The evolution of sex ratios and sex-determining systems. *Trends Ecol. Evol.* 22:292-97.

Umen, J. G. 2011. Evolution of sex and mating loci: An expanded view from Volvocine algae. *Curr. Opin. Microbiol.* 14:634-41.

Urushihara, H. & Muramoto, T. 2006. Genes involved in *Dictyostelium discoideum* sexual reproduction. *Eur. J. Cell Biol.* 85:961-68.

Vanormelingen, P., Vanelslander, B., Sato, S., Gillard, J., Trobajo, R., Sabbe, K. & Vyverman, W. 2013. Heterothallic sexual reproduction in the model diatom *Cylindrotheca*. *Eur. J. Phycol.* 48:93-105.

Vanstechelman, I. 2013. *Identification, characterization and evolution of the mating type locus in diatoms.* PhD, Ghent University.

Vanstechelman, I., Sabbe, K., Vyverman, W., Vanormelingen, P. & Vuylsteke, M. 2013. Linkage mapping identifies the sex determining region as a single locus in the pennate diatom *Seminavis robusta*. *PLoS ONE* 8:e60132.

Vaulot, D. & Chisholm, S. W. 1987. Flow cytometric analysis of spermatogenesis in the diatoms *Thalassiosira weissflogii* (Bacillariophiceae). *J. Phycol.* 23:132-37.

von Dassow, P., Chepurnov, V. A. & Armbrust, E. V. 2006. Relationships between growth rate, cell size, and induction of spermatogenesis in the centric diatom *Thalassiosira weissflogii* (Bacillariophyta). *J. Phycol.* 42:887-99.

von Dassow, P., Ogata, H., Probert, I., Wincker, P., Da Silva, C., Audic, S., Claverie, J. M. & de Vargas, C. 2009. Transcriptome analysis of functional differentiation between haploid and diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell. *Genome Biol.* 10:R114.

von Dassow, P. & Montresor, M. 2011. Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. *J. Plankton Res.* 33:3-12.

von Stosch, H. A. 1954. Die oogamie von *Biddulphia mobiliensis* und die bisher bekannten auxosporebildungen bei den centrales. *Rapport d'activité de la Commission du VIII Congrès International de Botanie.* pp. 58-68.

von Stosch, H. A. 1956. Entwicklungsgeschichtliche untersuchungen an zentrischen diatomeen. II. Geschlechtszellenreifung, befruchtung und auxosporenbildung einiger grundbewohnender biddulphiaceen der Nordsee. *Arch. Mikrobiol.* 23:327-65.

von Stosch, H. A. 1958. Entwicklungsgeschichtliche untersuchungen an zentrischen diatomeen. III. Die spermatogenese von *Melosira moniliformis* Agardh. *Arch. Mikrobiol.* 31:272–84.

von Stosch, H. A. & Drebes, G. 1964. Entwicklungsgeschichtliche untersuchungen an zentrischen diatomeen. IV. Die planktondiatomee *Stephanopyxis turris* - ihre behandlung und entwicklungsgeschichte. *Helgol. Wiss. Meeresunters.* 11:209-57.

von Stosch, H. A., Theil, G. & Kowallik, K. 1973. Entwicklungsgeschichtiliche untersuchungen an zentrischen diatomeen. V. Bau und lebenszyklus von *Chaetoceros didymum*, mit beobachtungen über einige andere arten der gattung. *Helgol. Wiss. Meeresunters.* 25:384-445.

Wang, Z., Gerstein, M. & Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Gen.* 10:57-63.

Whelan, S. & Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a Maximum-Likelihood approach. *Mol. Biol. Evol.* 18:691-99.

Yang, G., Ying, S., Yuanyuan, S., Linan, Z., Shanshan, G., Bingjun, L., Xiaojie, L., Zhiling, L., Yizhou, C., Yushan, Z. & Wenquan, W. 2009. Construction and characterization of a tentative amplified fragment length polymorphism–simple sequence repeat linkage map of *Laminaria* (Laminariales, Phaeophyta). *J. Phycol.* 45:873-78.

Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* 21:809-18.

Zanetti, S. & Puoti, A. 2013. Sex determination in the *Caenorhabditis elegans* germline. *In*: Schedl, T. [Ed.] *Germ Cell Development in C. elegans.* Springer New York, pp. 41-69.

Zhang, N., Zhang, L., Tao, Y., Guo, L., Sun, J., Li, X., Zhao, N., Peng, J., Li, X. & Zeng, L. 2015a. Construction of a high density SNP linkage map of kelp (*Saccharina japonica*) by sequencing Taq I site associated DNA and mapping of a sex determining locus. *BMC Genomics* 16:189.

Zhang, Q., Liu, C., Liu, Y., VanBuren, R., Yao, X., Zhong, C. & Huang, H. 2015b. High-density interspecific genetic maps of kiwifruit and the identification of sex-specific markers. *DNA Res.* 22:367-75.

APPENDIX A

List of differentially expressed genes between MT+ and MT- samples

Appendix A reports the list of the 91 significantly differentially expressed transcripts between mating types.

In the table are reported:

- Row.names of the transcripts
- logFC (logarithmic Fold Change)
- logCPM (logarithmic Counts Per Millions)
- PValue
- FDR (False Discovery Rate)
- CIIP, HCUN, HATT, CIIO, HCUH and HCUO (CPM for each RNA-seq library)
- QueryLength (nucleotidic length of the transcript)
- HSPNameSP, HSPLengthSP, HSPScoreSP, HITLengthSP, QCoverageSP, HCoverageSP, DescriptionSP (parameters of the predicted protein according to Swiss Prot)
- HSPNameUf, HSPLengthUf, HSPScoreUf, HITLengthUf, QCoverageUf, HCoverageUf, DescriptionUf (parameters of the predicted protein according to UniProt Reference Clusters)
- Taxonomy (organisms exhibiting the highest BLAST score identity with the query sequence)
- BPId, BPDesc (parameters of the biological process of the protein predicted according to GO terms)
- MFId, MFDesc (parameters of the molecular function of the protein predicted according to GO terms)
- CCId, CDName, CDStartEnd, CDScore, CDDesc (parameters of the cellular components of the protein predicted according to GO terms)

Row.names	logFC	logCPM	PValue	FDR	CIIP	HCUN	HATT	CIIO	HCUH	HCUO	Query Length	HSP Name SP	HSP Length SP	HSP ScoreSP	HIT Length SP	QCoverage SP	HCoverage SP	DescriptionSP	HSP NameUf	HSP LengthUf	HSP ScoreUf	HIT LengthUf	QCoverage Uf	HCoverage Uf	DescriptionUf	Taxonomy	BPId	BPDesc	MFId	MFDesc	CCId	CCDesc
comp13283_c0_seq1	-10,44	8,79	7,98E-08	1,88E-04	0,70	0,70	0,17	1068,26	1621,92	0,75	901																					
comp29861_c0_seq1	-3,61	5,79	1,28E-04	4,47E-02	2,66	13,76	9,31	144,86	151,75	15,74	2122	C0LGQ5	996	3,00E-23	1249	46,93685	22,33787	LRR receptor-like serine/threonine-protein kinase GSO1	UniRef90_K0R345	1359	3,00E-29	726	64,04336	61,01928	Uncharacterized protein (Fragment)	Thalassiosira oceanica						
comp21252_c0_seq1	-3,55	4,94	7,68E-05	2,98E-02	2,75	5,42	5,88	132,66	28,13	12,31	1209	Q9JLI4	765	8,00E-37	292	63,27543	85,27397	Elongation of very long chain fatty acids protein 2	UniRef90_Q5SE76	804	3,00E-141	272	66,50124	98,52941	Polyunsaturated fatty acid elongase 1	Thalassiosira pseudonana					GO:0016021	integral to membrane
comp29861_c0_seq2	-3,78	4,95	2,73E-05	1,25E-02	1,91	6,24	4,36	92,16	76,06	8,48	1170	P0DL10	819	3,00E-17	996	70	26,90763	Leucine-rich repeat receptor-like kinase protein THICK TASSEL DWARF1	UniRef90_B8C8C8	885	1,00E-22	315	75,64103	86,66667	Putative uncharacterized protein (Fragment)	Thalassiosira pseudonana						
comp26321_c0_seq1	-2,96	3,82	1,65E-04	5,49E-02	1,91	3,50	3,93	49,91	11,64	16,85	1193	Q9T0H9	666	2,00E-36	359	55,82565	62,67409	GDT1-like protein 2 chloroplastic	UniRef90_B7FZE2	636	2,00E-57	208	53,31098	100	Predicted protein (Fragment)	Phaeodactylum tricornutum (strain CCAP 1055/1)					GO:0016020	membrane
comp33825_c0_seq1	-3,03	4,00	2,03E-04	5,99E-02	2,10	6,36	1,24	47,43	21,65	20,05	463								UniRef90_B7G073	318	3,00E-14	237	68,68251	41,35021	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)					GO:0016020	membrane
comp22054_c0_seq1	-2,69	4,32	3,52E-04	9,08E-02	1,96	10,39	3,73	45,14	40,61	22,65	1221	P80030	897	2,00E-132	385	73,46437	75,84416	Enoyl- acyl-carrier-protein reductase NADH chloroplastic	UniRef90_B7FS72	903	6,00E-176	310	73,95577	97,09677	Enoyl-acp reductase	Phaeodactylum tricornutum (strain CCAP 1055/1)	GO:0006633	fatty acid biosynthetic process	GO:0004318	enoyl-[acyl-carrier-protein] reductase (NADH) activity		
comp17221_c0_seq1	-3,17	3,68	1,76E-04	5,69E-02	1,54	4,47	1,18	43,23	13,64	15,02	364	Q78424	174	9,00E-12	291	47,8022	64,83516	Uncharacterized protein ycf19	UniRef90_B7G073	270	6,00E-27	237	74,17582	40,50633	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)					GO:0016020	membrane
comp26707_c0_seq1	-5,15	3,06	9,03E-06	6,03E-03	0,23	0,57	0,53	42,90	2,91	3,45	1776	G5EGA5	1026	4,00E-38	376	57,77027	84,84043	Delta(12) fatty acid desaturase fat-2	UniRef90_B7GEK2	1248	1,00E-160	435	70,27027	94,25287	Precursor of desaturase omega-3 desaturase	Phaeodactylum tricornutum (strain CCAP 1055/1)	GO:0006629	lipid metabolic process				
comp22580_c1_seq1	-2,93	4,84	1,92E-04	5,90E-02	1,03	12,13	8,04	36,09	55,63	72,70	891								UniRef90_B7FVU8	849	2,00E-69	308	95,2862	93,18182	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)						
comp20210_c0_seq1	-3,79	4,17	2,65E-05	1,25E-02	3,54	0,55	1,41	33,11	40,43	33,21	1254								UniRef90_B8BTY4	807	2,00E-83	341	64,35407	85,92375	Predicted protein	Thalassiosira pseudonana						
comp28121_c0_seq1	-2,97	3,47	3,39E-04	8,82E-02	1,17	5,10	0,96	24,29	20,78	17,11	1175								UniRef90_B7FNV7	643	3,00E-31	188	46,21277	96,80851	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)						
comp24462_c0_seq2	-13,28	2,61	1,90E-14	1,94E-10	0,00	0,00	0,00	22,89	7,79	7,13	2649								UniRef90_F8B226	480	2,00E-06	662	18,12005	22,96073	Uncharacterized protein	Frankia symbiont subsp. Datisca glomerata			GO:0016787	hydrolase activity		
comp32905_c0_seq1	-7,04	3,58	8,37E-06	5,71E-03	0,37	0,00	0,00	20,68	23,78	32,01	723	Q7RTY7	309	3,00E-07	1134	42,73859	8,641975	Ovochymase-1	UniRef90_Q7Z0G5	309	1,00E-09	262	42,73859	36,25954	Chymotrypsin	Phlebotomus papatasi	GO:0006508	proteolysis	GO:0004252	serine-type endopeptidase activity		
comp22340_c0_seq2	-3,20	3,16	1,56E-04	5,34E-02	0,84	0,78	3,76	20,24	19,05	11,19	1935	Q9R0L1	234	8,00E-08	492	12,09302	17,88618	Heat shock factor protein 4	UniRef90_B7S408	285	1,00E-18	378	14,72868	25,39683	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)			GO:0043565]---[GO:0003700	sequence-specific DNA binding]---[sequence-specific DNA binding transcription factor activity	GO:0005634	nucleus
comp22480_c0_seq3	-11,63	3,71	6,81E-22	2,09E-17	0,00	0,02	0,00	19,97	33,43	30,57	3013								UniRef90_B8LCX4	2913	0	1177	96,68105	83,94223	Predicted protein	Thalassiosira pseudonana					GO:0016021	integral to membrane
comp22340_c0_seq1	-3,18	3,14	1,73E-04	5,65E-02	0,84	0,78	3,76	19,61	18,95	11,11	1793	Q9R0L1	234	7,00E-08	492	13,05075	17,88618	Heat shock factor protein 4	UniRef90_B7S408	285	1,00E-18	378	15,89515	25,39683	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)			GO:0043565]---[GO:0003700	sequence-specific DNA binding]---[sequence-specific DNA binding transcription factor activity	GO:0005634	nucleus
comp31407_c0_seq11	-3,19	3,31	9,65E-05	3,61E-02	0,89	3,77	1,01	14,95	19,43	22,90	1185								UniRef90_K0SE47	330	4,00E-13	653	27,8481	17,76417	Uncharacterized protein	Thalassiosira oceanica						
comp5S770_c0_seq1	-12,88	2,22	8,26E-07	9,39E-04	0,00	0,00	0,00	14,49	13,33	0,00	245																	GO:0009055]-[GO:0051536]-[GO:0046872	electron carrier activity]---[iron-sulfur cluster binding]---[metal ion binding		
comp31481_c0_seq1	-2,85	3,68	3,76E-04	9,46E-02	2,47	4,63	1,25	12,01	23,82	38,13	1623								UniRef90_B8CDS1	807	7,00E-48	334	49,72274	83,53293	Predicted protein	Thalassiosira pseudonana						
comp29236_c0_seq1	-4,12	1,57	3,09E-04	8,31E-02	0,14	0,65	0,12	10,20	4,01	2,90	1918								UniRef90_K0RA30	1515	2,00E-46	645	78,98853	83,10078	Uncharacterized protein	Thalassiosira oceanica			GO:0005247	voltage-gated chloride channel activity	GO:0016020	membrane
comp20279_c0_seq5	-12,20	1,55	3,70E-06	3,16E-03	0,00	0,00	0,00	9,99	7,35	0,00	1613																					
comp20279_c0_seq4	-12,17	1,52	3,91E-06	3,20E-03	0,00	0,00	0,00	9,89	7,14	0,00	1456																					
comp32904_c0_seq1	-6,48	2,59	4,77E-05	2,02E-02	0,28	0,00	0,00	9,22	11,93	17,17	544	P80646	393	1,00E-10	245	72,24265	52,2449	Chymotrypsin B	UniRef90_H9L027	402	9,00E-14	252	73,89706	51,19048	Uncharacterized protein (Fragment)	Gallus gallus	GO:0006508	proteolysis	GO:0004252	serine-type endopeptidase activity		
comp29283_c0_seq1	4,17	6,27	2,24E-04	6,41E-02	246,93	10,49	47,67	8,94	6,19	11,49	1742	Q5BGA7	876	2,00E-93	819	50,28703	92,47649	Probable NAD(P)H-dependent D-xylose reductase xyl1	UniRef90_L0DCZ5	921	7,00E-112	826	52,87026	96,01227	Aldo/keto reductase diketogulonate reductase	Singulisphaera acidiphila (strain ATCC BAA-1392 / DSM 18658 / VKM B-2454 / MOB10)			GO:0016491	oxidoreductase activity		
comp32405_c0_seq8	2,80	4,81	2,70E-05	1,25E-02	28,45	37,50	79,36	8,82	5,81	7,94	1715	O54965	153	6,00E-08	381	8,921283	12,86089	E3 ubiquitin-protein ligase RNF13	UniRef90_K0TCR7	162	6,00E-09	385	9,446064	13,76623	Uncharacterized protein	Thalassiosira oceanica			GO:0008270	zinc ion binding		
comp16327_c0_seq1	-4,87	2,93	7,32E-07	8,65E-04	0,56	0,13	0,62	8,69	13,86	25,56	1551								UniRef90_R1DHG5	813	4,00E-45	488	52,41779	55,32787	Uncharacterized protein	Emiliania huxleyi CCMP1516						
comp32405_c0_seq4	2,82	4,80	2,27E-05	1,16E-02	27,89	37,63	79,43	8,40	5,83	7,98	1576	O54965	153	5,00E-08	381	9,708122	12,86089	E3 ubiquitin-protein ligase RNF13	UniRef90_K0TCR7	162	5,00E-09	385	10,27919	13,76623	Uncharacterized protein	Thalassiosira oceanica			GO:0008270	zinc ion binding		
comp32405_c0_seq7	2,79	4,60	2,49E-05	1,23E-02	23,08	33,77	70,43	7,31	5,28	7,17	1855	O54965	156	7,00E-08	381	8,409704	13,12336	E3 ubiquitin-protein ligase RNF13	UniRef90_K0TCR7	162	4,00E-09	385	8,733154	13,76623	Uncharacterized protein	Thalassiosira oceanica			GO:0008270	zinc ion binding		
comp32405_c0_seq2	2,81	4,59	2,14E-05	1,13E-02	22,52	33,90	70,51	6,89	5,30	7,20	1716	O54965	156	6,00E-08	381	9,090909	13,12336	E3 ubiquitin-protein ligase RNF13	UniRef90_K0TCR7	162	4,00E-09	385	9,440559	13,76623	Uncharacterized protein	Thalassiosira oceanica			GO:0008270	zinc ion binding		
comp22428_c0_seq4	3,02	4,36	1,91E-04	5,90E-02	30,36	8,05	65,31	6,43	3,46	4,66	896	Q9LDY2	459	9,00E-75	358	51,22768	42,4581	2-oxoisovalerate dehydrogenase subunit beta 2 mitochondrial	UniRef90_B7FUY5	459	1,00E-85	323	51,22768	47,05882	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)			GO:0003824	catalytic activity		
comp21130_c0_seq1	2,57	3,94	2,65E-04	7,24E-02	26,63	20,41	21,16	4,93	3,68	5,82	3580	Q0R4F1	462	5,00E-08	635	12,90503	20,62992	ATP-dependent DNA helicase PIF1	UniRef90_F4PS42	636	1,00E-16	1554	17,76536	12,41956	Putative uncharacterized protein	Dictyostelium fasciculatum (strain SH3)			GO:0005509	calcium ion binding		
comp27491_c0_seq3	3,93	5,74	3,12E-04	8,33E-02	168,21	11,71	29,66	4,39	5,15	12,44	1409	O34371	531	7,00E-07	428	37,6863	41,1215	Putative oxidoreductase YteT	UniRef90_B8BUH6	1149	1,00E-134	413	81,5472	93,7046	Predicted protein	Thalassiosira pseudonana			GO:0016491	oxidoreductase activity		

comp27491_c0_seq1	4,00	5,66	2,47E-04	6,90E-02	159,35	11,14	28,46	4,12	4,56	11,17	1278	O34371	531	6,00E-07	428	41,5493	41,1215	Putative oxidoreductase YteT	UniRef90_B8BUH6	1143	1,00E-134	413	89,43662	93,22034	Predicted protein	Thalassiosira pseudonana			GO:0016491	oxidoreductase activity		
comp29120_c0_seq2	3,09	4,67	4,27E-06	3,28E-03	24,16	43,27	68,58	3,84	5,17	8,55	2783								UniRef90_K0SPX5	894	0	640	32,12361	45,9375	Uncharacterized protein	Thalassiosira oceanica			GO:0004683	calmodulin-dependent protein kinase activity		
comp22480_c0_seq2	-12,12	1,48	1,36E-10	3,79E-07	0,00	0,00	0,00	3,04	7,44	7,05	1101								UniRef90_B8LCX4	1017	6,00E-38	1177	92,37057	29,9915	Predicted protein	Thalassiosira pseudonana					GO:0016021	integral to membrane
comp25742_c0_seq1	-7,63	0,51	1,17E-05	7,48E-03	0,05	0,00	0,00	3,00	2,58	3,08	899								UniRef90_C5H670	813	2,00E-08	314	90,43382	79,29936	Rapid-growth-like protein 14	Skeletonema costatum						
comp25269_c0_seq1	4,35	4,80	3,34E-07	5,40E-04	67,71	14,45	47,28	2,98	2,17	3,33	689																					
comp25269_c0_seq2	4,51	4,87	1,68E-07	3,23E-04	71,77	15,40	49,38	2,83	2,05	3,16	746																					
comp24462_c0_seq1	-9,93	-0,65	1,38E-05	8,49E-03	0,00	0,00	0,00	2,56	0,57	0,56	327																					
comp25256_c0_seq1	-7,59	0,51	3,31E-07	5,40E-04	0,00	0,02	0,02	2,54	1,46	4,96	1916								UniRef90_A9SAS7	1047	2,00E-26	449	54,64509	77,50557	Predicted protein	Physcomitrella patens subsp. patens						
comp17886_c0_seq1	-7,14	0,71	1,64E-06	1,63E-03	0,00	0,06	0,00	2,27	3,87	3,91	465																					
comp8253_c0_seq1	-7,20	-0,82	1,83E-04	5,84E-02	0,00	0,00	0,02	2,25	0,65	0,34	247																					
comp32211_c0_seq2	2,77	4,06	1,96E-04	5,93E-02	24,30	26,48	28,70	1,85	3,72	8,81	3180																					
comp32211_c0_seq2	2,77	4,06	1,96E-04	5,93E-02	24,30	26,48	28,70	1,85	3,72	8,81	3180																					
comp31002_c1_seq2	-5,71	-0,87	3,25E-04	8,59E-02	0,00	0,02	0,03	1,81	0,36	0,99	776																					
comp31002_c1_seq2	-5,71	-0,87	3,25E-04	8,59E-02	0,00	0,02	0,03	1,81	0,36	0,99	776																					
comp12430_c1_seq1	-10,22	-0,36	1,75E-06	1,64E-03	0,00	0,00	0,00	1,72	1,37	1,57	1230																					
comp12430_c1_seq2	-9,97	-0,59	3,97E-06	3,20E-03	0,00	0,00	0,00	1,26	1,18	1,51	933																					
comp27022_c0_seq1	5,50	4,44	1,91E-05	1,05E-02	71,35	0,89	18,31	1,20	0,47	1,45	1001																					
comp7433_c0_seq1	4,25	3,49	1,97E-04	5,93E-02	36,56	6,68	2,13	0,99	1,94	0,71	1252																					
comp22480_c0_seq1	-10,44	-0,14	1,12E-06	1,18E-03	0,00	0,00	0,00	0,84	2,68	1,93	537								UniRef90_B8LCX4	372	4,00E-11	1177	69,27374	11,97961	Predicted protein	Thalassiosira pseudonana					GO:0016021	integral to membrane
comp24306_c0_seq1	5,65	2,89	9,11E-05	3,45E-02	3,82	29,75	11,01	0,82	0,06	0,00	1025																					
comp17312_c0_seq2	4,68	2,69	8,34E-05	3,20E-02	20,75	4,36	2,44	0,71	0,68	0,18	1060								UniRef90_UPI0002246CD6	330	6,00E-06	1112	31,13208	8,723022	UPI0002246CD6 related cluster	unknown						
comp24306_c0_seq2	6,11	2,94	1,35E-05	8,47E-03	4,48	30,25	11,51	0,61	0,06	0,00	1067																					
comp9495_c0_seq1	-5,84	0,71	3,67E-04	9,39E-02	0,00	0,15	0,00	0,61	6,42	2,72	201																					
comp16576_c0_seq2	4,65	2,37	2,14E-04	6,20E-02	1,77	1,07	30,42	0,59	0,13	0,54	592								UniRef90_B7G4T6	261	1,00E-10	1014	44,08784	8,678501	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)	GO:0006054	glucuronate catabolic process	GO:0008880	glucuronate isomerase activity		
comp6601_c0_seq1	8,61	6,12	5,55E-07	7,41E-04	23,46	398,26	0,96	0,57	0,44	0,08	1029																					
comp32596_c0_seq1	6,69	4,05	4,65E-06	3,48E-03	4,15	96,22	0,63	0,50	0,34	0,14	386																					
comp23156_c0_seq2	4,67	3,23	1,57E-04	5,34E-02	32,50	1,24	4,08	0,48	1,08	0,79	1865	P10961	348	4,00E-15	833	18,65952	15,006	Heat shock factor protein	UniRef90_B7G6V6	648	3,00E-72	454	34,74531	49,33921	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)			GO:0043565]---[GO:0003700	sequence-specific DNA binding]---[sequence-specific DNA binding transcription factor activity	GO:0005634	nucleus
comp17312_c0_seq3	4,66	2,25	1,85E-04	5,84E-02	15,72	3,14	1,72	0,48	0,59	0,12	683																					
comp6989_c0_seq1	4,81	2,13	1,09E-04	3,89E-02	7,37	0,34	16,58	0,40	0,25	0,34	907																					
comp7488_c0_seq1	5,27	4,24	7,37E-06	5,14E-03	3,17	112,11	0,91	0,40	0,47	0,69	1117								UniRef90_B5Y5H3	801	3,00E-60	715	71,70994	35,8042	Predicted protein	Phaeodactylum tricornutum (strain CCAP 1055/1)						
comp23886_c1_seq2	7,07	3,67	2,54E-13	1,30E-09	17,49	17,78	38,06	0,36	0,15	0,08	4827	Q9SMT7	1566	4,00E-33	514	32,44251	96,3035	4-coumarate--CoA ligase-like 10	UniRef90_S3N1F2	4191	2,00E-140	1433	86,82411	97,83671	Uncharacterized protein	Acinetobacter rudis CIP 110305						
comp23886_c1_seq1	7,07	3,67	2,54E-13	1,30E-09	17,49	17,78	38,06	0,36	0,15	0,08	4913	Q9SMT7	1566	5,00E-33	514	31,87462	96,3035	4-coumarate--CoA ligase-like 10	UniRef90_S3N1F2	4191	3,00E-140	1433	85,30429	97,83671	Uncharacterized protein	Acinetobacter rudis CIP 110305						
comp22837_c0_seq1	5,18	3,08	2,39E-05	1,20E-02	7,55	1,37	43,27	0,36	0,09	1,09	1489	P29787	669	1,00E-37	266	44,92948	84,58647	Trypsin 5G1	UniRef90_K0ST48	684	1,00E-40	541	45,93687	44,91682	Uncharacterized protein	Thalassiosira oceanica	GO:0006508	proteolysis	GO:0004252	serine-type endopeptidase activity		
comp25070_c0_seq2	-9,28	-1,24	4,87E-05	2,02E-02	0,00	0,00	0,00	0,27	1,27	0,89	265																					
comp20880_c0_seq1	6,04	1,88	3,99E-04	9,95E-02	14,83	0,08	0,67	0,21	0,06	0,12	274																					
comp11747_c0_seq1	-6,32	1,10	4,87E-05	2,02E-02	0,05	0,08	0,02	0,17	7,81	5,28	1189								UniRef90_D0NOR4	696	4,00E-23	322	58,53659	72,98137	Putative uncharacterized protein	Phytophthora infestans (strain T30-4)			GO:0005525	GTP binding		
comp26595_c0_seq2	9,60	5,43	1,45E-06	1,48E-03	42,34	202,57	0,12	0,13	0,21	0,00	1925	Q9LP24	474	6,00E-17	1120	24,62338	13,92857	Probable leucine-rich repeat receptor-like protein kinase At1g35710	UniRef90_D7FL98	372	2,00E-19	1282	19,32468	9,594384	LRR-GTPase of the ROCO family	Ectocarpus siliculosus	GO:0007264	small GTPase mediated signal transduction	GO:0005525	GTP binding		
comp26595_c0_seq1	9,76	5,42	8,68E-07	9,52E-04	40,76	202,51	0,14	0,13	0,17	0,00	2022	Q9LP24	474	6,00E-17	1120	23,44214	13,92857	Probable leucine-rich repeat receptor-like protein kinase At1g35710	UniRef90_D7FL98	372	2,00E-19	1282	18,39763	9,594384	LRR-GTPase of the ROCO family	Ectocarpus siliculosus	GO:0007264	small GTPase mediated signal transduction	GO:0005525	GTP binding		
comp26595_c0_seq3	9,53	5,49	3,58E-07	5,50E-04	44,35	211,27	0,14	0,10	0,25	0,02	1804	Q9SSL9	954	4,00E-31	1123	52,88248	30,09795	Leucine-rich repeat receptor-like protein kinase PEPR1	UniRef90_Q9ZTK1	930	1,00E-30	968	51,55211	32,1281	Disease resistance protein	Solanum lycopersicum						
comp26595_c0_seq4	9,67	5,48	2,10E-07	3,79E-04	42,76	211,21	0,15	0,10	0,21	0,02	1901	Q9LVP0	708	9,00E-18	1102	37,24356	21,05263	Probable leucine-rich repeat receptor-like protein kinase At5g63930	UniRef90_D8QU71	723	5,00E-22	1078	38,03261	24,39703	Putative uncharacterized protein	Selaginella moellendorffii			GO:0005524]---[GO:0004674	ATP binding]---[protein serine/threonine kinase activity		
comp49483_c0_seq1	5,54	0,93	2,26E-04	6,41E-02	0,23	0,57	11,58	0,10	0,04	0,10	317																					

comp23861_c0_seq1	6,52	0,33	1,52E-05	9,12E-03	1,21	1,85	4,49	0,06	0,00	0,02	963																				
comp6440_c0_seq1	7,75	2,33	1,20E-04	4,24E-02	0,84	29,96	0,00	0,06	0,06	0,02	1502								UniRef90_K8Z515	981	5,00E-46	764	65,31292	35,07853	Uncharacterized protein	Nannochloropsis gaditana CCMP526						
comp23284_c0_seq1	9,03	3,63	1,01E-04	3,70E-02	0,75	76,33	0,00	0,06	0,08	0,00	512																				
comp21038_c0_seq1	9,52	4,11	2,58E-05	1,24E-02	0,70	107,66	0,02	0,06	0,08	0,00	452																				
comp33110_c0_seq1	-8,68	-1,79	3,32E-04	8,72E-02	0,00	0,00	0,00	0,06	0,97	0,58	896																				
comp33193_c0_seq1	11,17	3,10	2,30E-15	3,54E-11	21,13	6,30	15,84	0,02	0,00	0,00	495																				
comp24201_c0_seq2	6,97	-0,61	1,04E-04	3,75E-02	2,56	0,48	0,74	0,02	0,00	0,02	502																				
comp31789_c0_seq1	6,00	-0,04	5,57E-05	2,19E-02	1,31	3,52	0,71	0,02	0,02	0,06	255																				
comp12331_c0_seq1	8,84	3,68	2,59E-06	2,34E-03	0,65	79,55	0,14	0,02	0,08	0,08	1742	O93428	1011	2,00E-63	396	58,03674	83,08081	Cathepsin D	UniRef90_B7GBN2	1032	9,00E-98	344	59,24225	97,09302	Predicted protein (Fragment)	Phaeodactylum tricornutum (strain CCAP 1055/1)	GO:0006508	proteolysis	GO:0004190	aspartic-type endopeptidase activity		
comp15702_c0_seq1	6,14	0,27	1,87E-04	5,84E-02	0,09	0,65	6,94	0,02	0,02	0,06	504								UniRef90_B8BU16	273	2,00E-19	317	54,16667	28,70662	Predicted protein	Thalassiosira pseudonana	GO:0032259	methylation	GO:0008168	methyltransferase activity		
comp6261_c0_seq1	12,64	1,94	2,47E-13	1,30E-09	4,76	5,33	12,82	0,00	0,00	0,00	1227								UniRef90_K0R0D8	519	1,00E-82	186	42,29829	98,92473	Uncharacterized protein	Thalassiosira oceanica						
comp28108_c0_seq2	10,55	2,48	4,62E-07	6,44E-04	3,96	28,57	0,05	0,00	0,02	0,00	1705	P38531	312	6,00E-134	467	18,29912	21,84154	Heat shock factor protein 3	UniRef90_M5WVD4	327	2,00E-13	195	19,17889	52,30769	Uncharacterized protein	Prunus persica			GO:0043565]---[GO:0003700	sequence-specific DNA binding]---[sequence-specific DNA binding transcription factor activity	GO:0005634	nucleus
comp28108_c0_seq1	10,43	2,36	6,25E-07	7,99E-04	3,50	26,67	0,05	0,00	0,02	0,00	1546	P38531	312	5,00E-134	467	20,18111	21,84154	Heat shock factor protein 3	UniRef90_M5WVD4	327	1,00E-13	195	21,15136	52,30769	Uncharacterized protein	Prunus persica			GO:0043565]---[GO:0003700	sequence-specific DNA binding]---[sequence-specific DNA binding transcription factor activity	GO:0005634	nucleus
comp23284_c0_seq2	10,50	3,31	4,15E-06	3,26E-03	1,96	59,22	0,02	0,00	0,04	0,00	940																				
comp33676_c0_seq1	10,02	-0,74	5,84E-06	4,17E-03	0,89	1,83	0,79	0,00	0,00	0,00	394																				
comp10158_c0_seq1	9,12	-1,90	2,00E-04	5,97E-02	0,84	0,08	0,76	0,00	0,00	0,00	258																				
comp23569_c0_seq1	6,60	-0,03	2,67E-04	7,24E-02	0,09	1,77	4,19	0,00	0,00	0,06	280																				

APPENDIX B

REST analyses of the genes resulted to be not differentially expressed between MT+ and MT- samples

REST analyses of the nine transcripts (see Chapter 2, section 2.3.5) (Table B1) resulted not differentially expressed according to MT or not differentially expressed at all. Graphical representation of their expression ratio is reported in Figs. B1, B2, B3, B4, B5, B6, B7, B8 and B9.

Table B1: List of the transcripts resulted not differentially expressed afted qRT-PCR validation. Normalized counts provided for S1+ = Sy373 small, S2+ = B856 small, L2+ = B856 large, S1- = Sy379 small, S2- = B857 small, L2- = B857 large. LogFC= 2log fold change, Pvalue = p-value and FDR= False discovery rate.

Transcript ID	**logFC**	**PValue**	**FDR**	**S1-**	**S2-**	**L2-**	**S1+**	**S2+**	**L2+**
comp24462_c0_seq2	-13.28	1.90E-14	1.94E-10	0.00	0.00	0.00	22.89	7.79	7.13
comp22480_c0_seq3	-11.63	6.81E-22	2.09E-17	0.00	0.02	0.00	19.97	33.43	30.57
comp23156_c0_seq2	4.67	1.57E-04	5.34E-02	32.50	1.24	4.08	0.48	1.08	0.79
comp6261_c0_seq1	12.64	2.47E-13	1.30E-09	4.76	5.33	12.82	0.00	0.00	0.00
comp27491_c0_seq3	3.93	3.12E-04	8.33E-02	168.2	11.71	29.66	4.39	5.15	12.44
comp27022_c0_seq1	5.50	1.91E-05	1.05E-02	71.35	0.89	18.31	1.20	0.47	1.45
comp25269_c0_seq1	4.35	3.34E-07	5.40E-04	67.71	14.45	47.28	2.98	2.17	3.33
comp29120_c0_seq2	3.09	4.27E-06	3.28E-03	24.16	43.27	68.58	3.84	5.17	8.55
comp31481_c0_seq1	-2.85	3.76E-04	9.46E-02	2.47	4.63	1.25	12.01	23.82	38.13

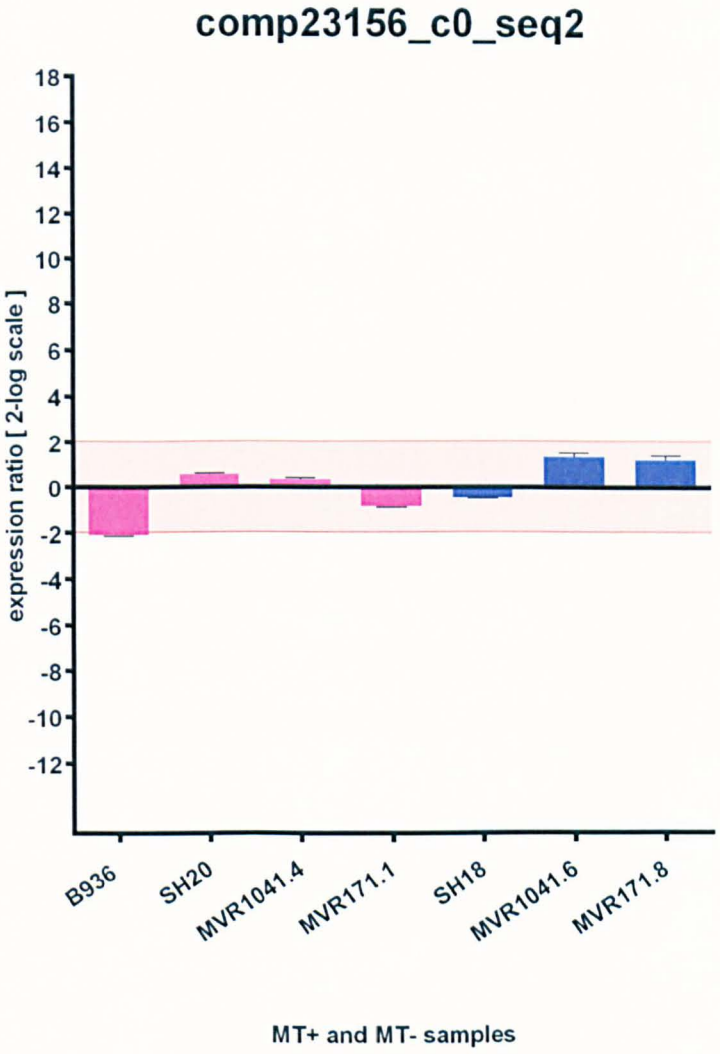


Figure B1: REST analysis of comp23156_c0_seq2. Reference condition: B935 MT+, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

comp23156_c0_seq2 resulted not differentially expressed in both the MTs.

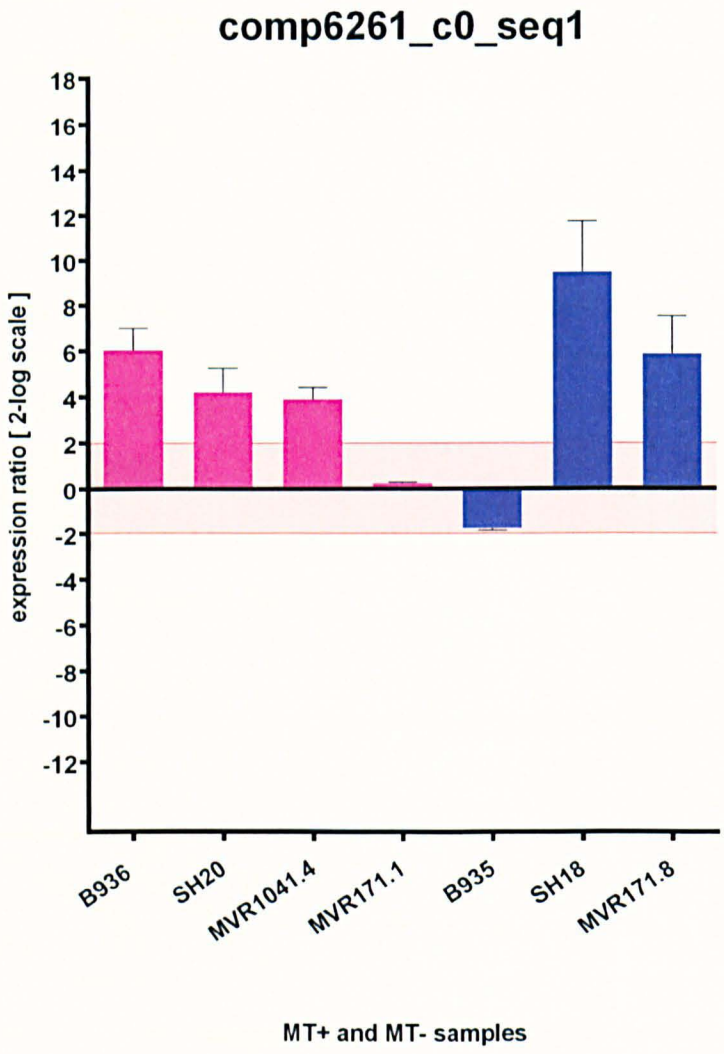


Figure B2: REST analysis of comp6261_c0_seq1. Reference condition: MVR1041.4 MT-, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

Comp6261_c0_seq1 was expected to be overexpressed in MT-. It was annotated as an uncharacterized protein with a conserved domain of S-adenosylmethionine-dependent methyltransferases in the principal ORF (RF+2). The RT-PCR and qRT-PCR gave discordant results with respect to the *in silico* prediction and its expression varied among the MT- strains.

Referring at the differential expression analysis made for both *P. multistriata* transcriptome and the 'sensing transcriptome', comp6261_c0_seq1 resulted differentially expressed in different samples but with no correlation to time or sexualisation (data not shown). In the

sensing transcriptome and in *P.multistriata* transcriptome comp6261_c0_seq1 resulted missing in sample B856 probably causing the misleading result of the differential expression analysis. However B856 was amplified on gDNA by PCR, giving positive results.

Transcripts comp24462_c0_seq2 and comp22480_c0_seq3 were excluded from the REST analyses because in some samples their CT values resulted to be 'Undetermined' from the qRT-PCR data, meaning that the transcript was so little expressed that it was impossible to be detected. However, their differential expression rates resulted not related to mating types. Please, about Figs. B3 and B4, remind that the numerical value of the CT is inversely related to the amount of amplicon in the reaction, i.e., the lower the CT, the larger the amount of amplicon.

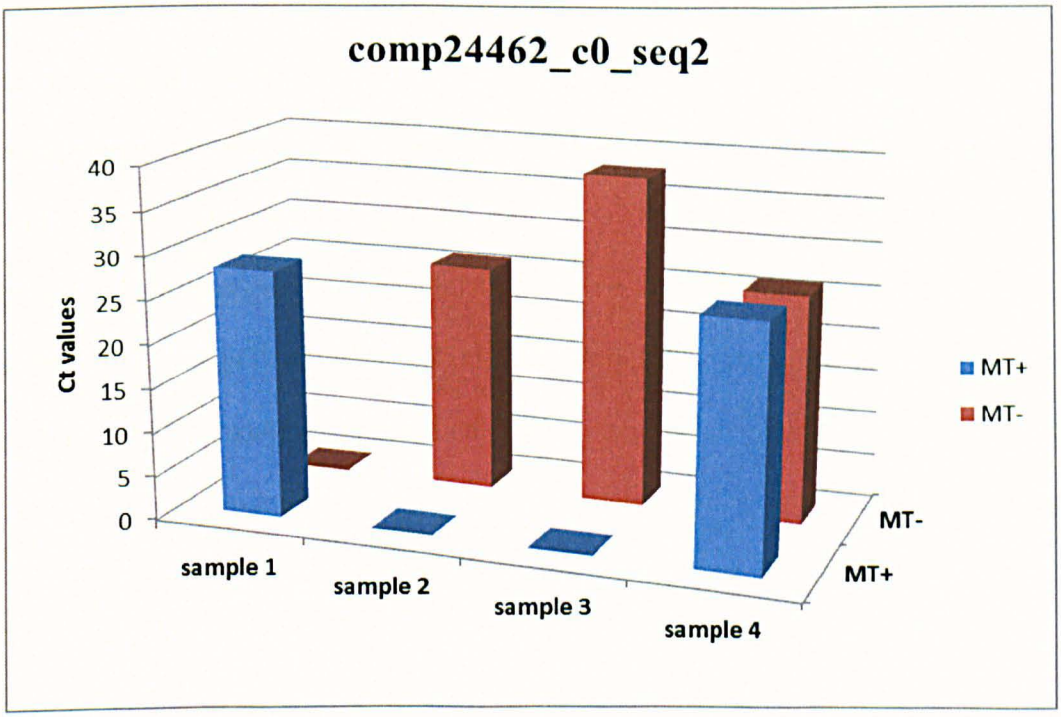


Figure B3: Graph reporting the Ct values of 4 MT+ samples in blue and 4MT- samples in red for comp24462_c0_seq2. Where the bars are unrepresented it means that the samples was 'Undetermined' so no Ct value was detected by the qPCR.

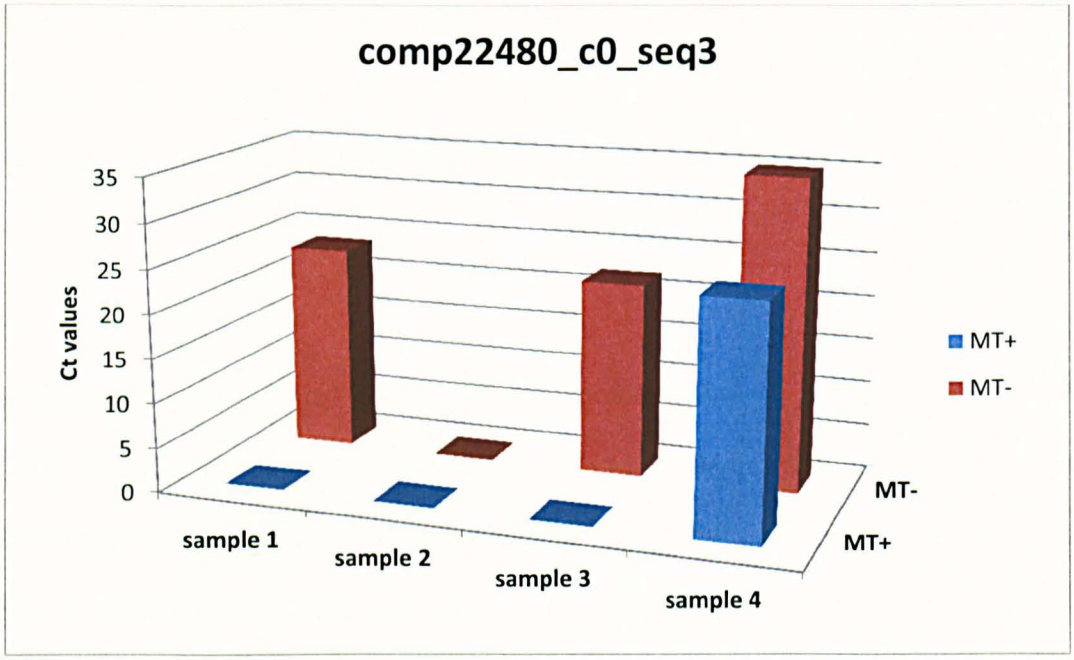


Figure B4: Graph reporting the Ct values of 4 MT+ samples in blue and 4 MT- samples in red for comp22480_c0_seq3. The zero bars represent the 'Undetermined' samples.

comp22480_c0_seq3 was one of the best results obtained by the differential expression analysis. Expected to be overexpressed in MT+, it had an EamA-like transporter family conserved domain in the principal ORF (RF+3). Also in the sensing transcriptome it looked to be more expressed in MT+ samples, with no relevant correlation to time or sexualisation, while in the MT- it was completely absent. However it did not result significantly differentially expressed in the qRT-PCR validations (Fig.B4). This data could be explained by the very low level of expression and further analyses will be considered to confirm this result.

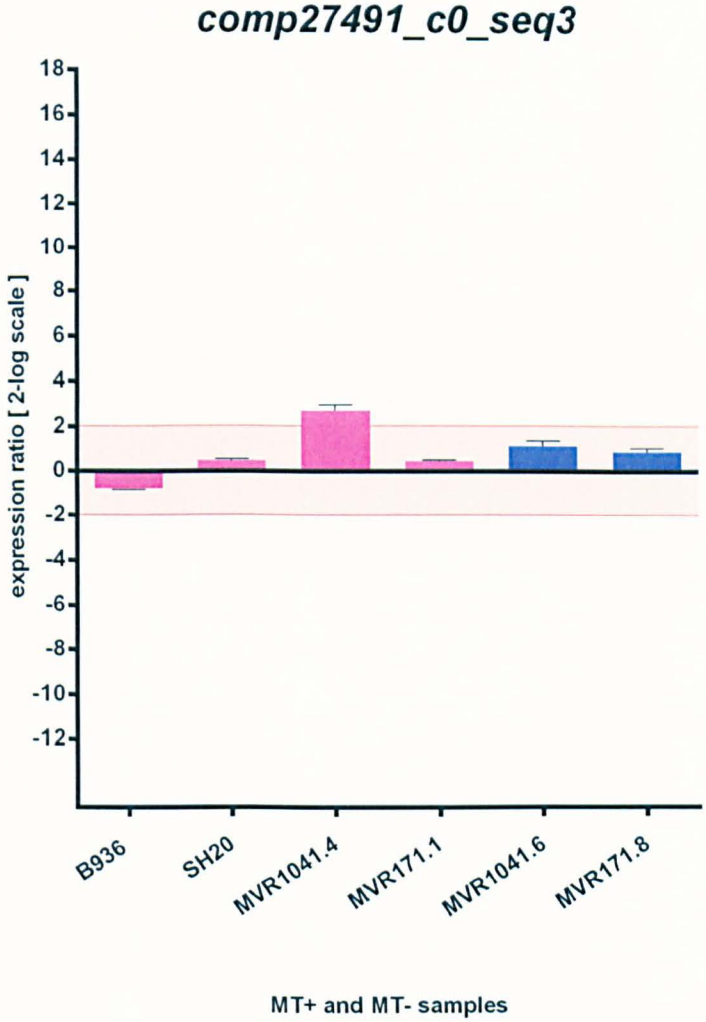


Figure B5: REST analysis of comp27491_c0_seq3. Reference condition: SH18+, reference genes: *CDK, TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

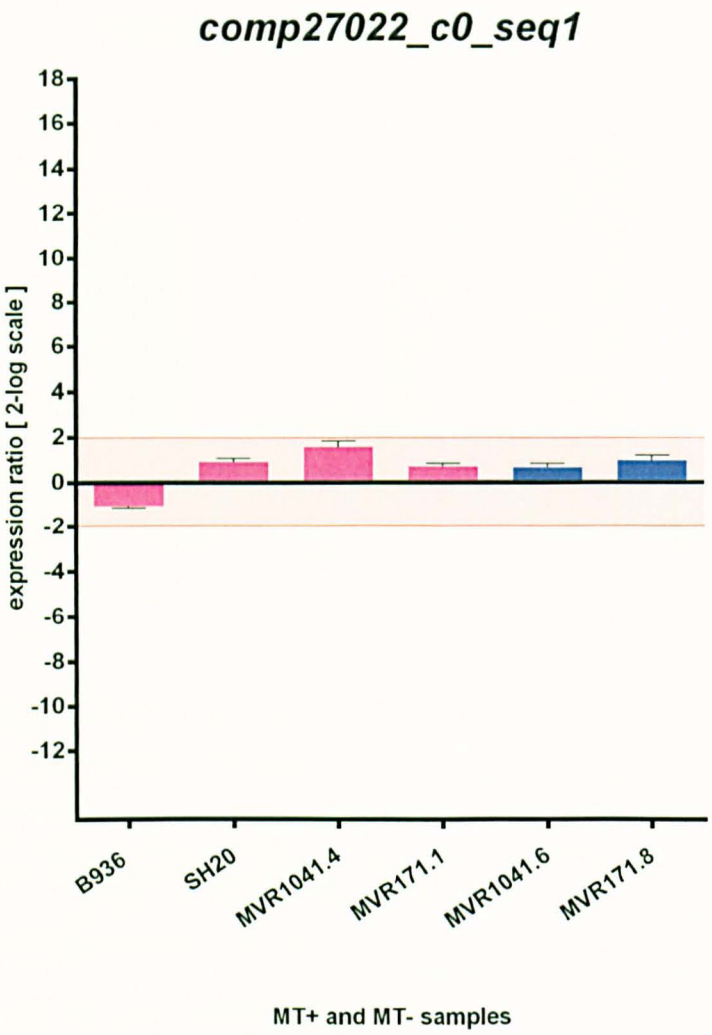


Figure B6: REST analysis of comp27022_c0_seq1. Reference condition: SH18+, reference genes: *CDK*, *TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

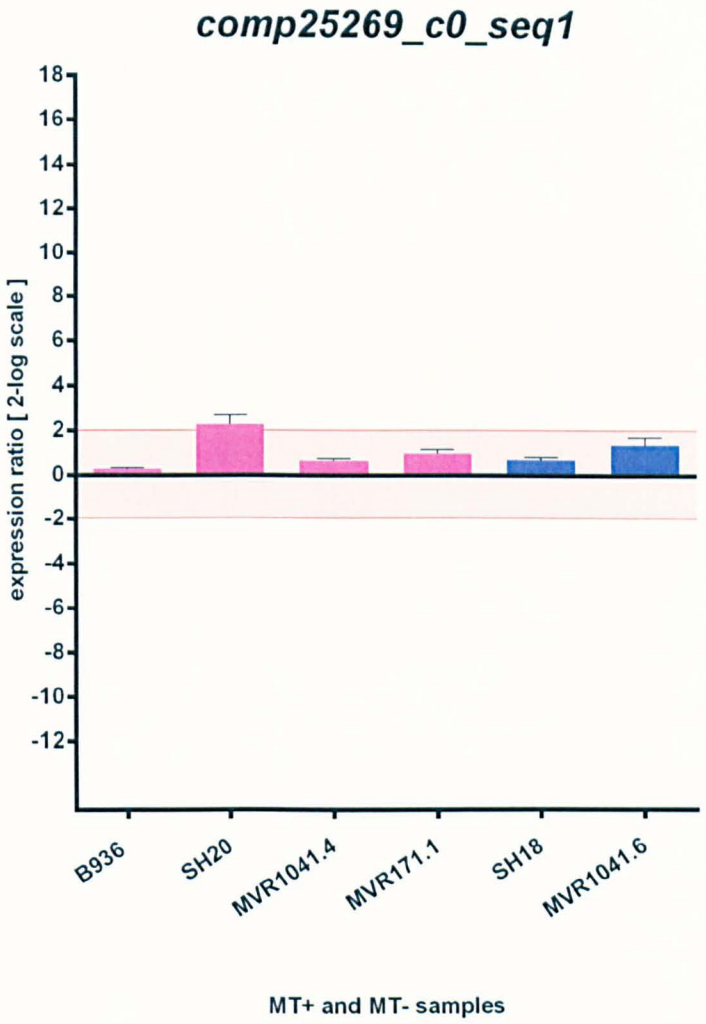


Figure B7: REST analysis of comp25269_c0_seq1. Reference condition: MVR171.8+, reference genes: *CDK, TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

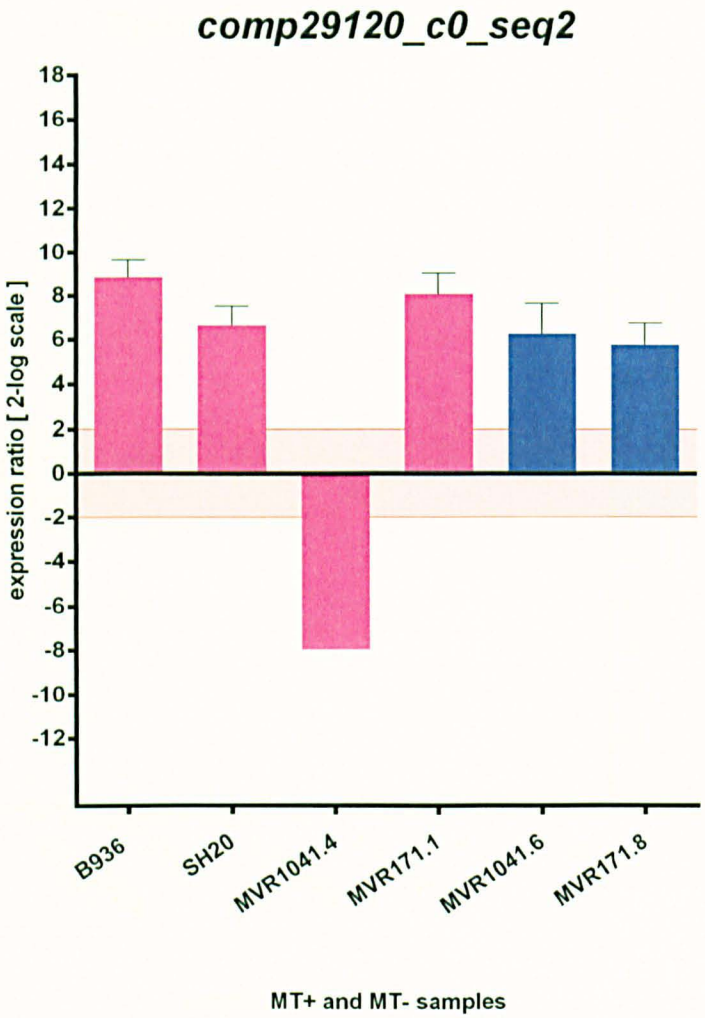


Figure B8: REST analysis of comp29120_c0_seq2. Reference condition: SH18+, reference genes: *CDK, TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples.

comp29120_c0_seq, although being differentially expressed, it was not in relation to MT.

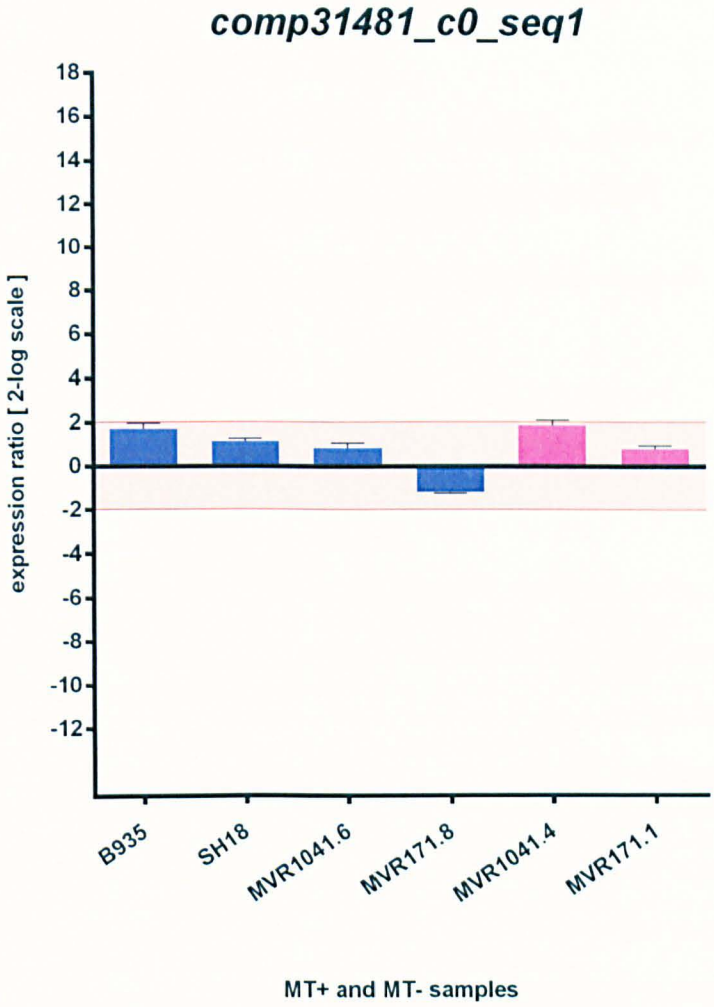


Figure B9: REST analysis of comp31481_c0_seq. Reference condition: SH20-, reference genes: *CDK, TUB A* and *TUB B*. Blue bars: the expression of MT+ samples, pink bars: the expression of MT- samples

.

APPENDIX C

Protein multiple sequences alignments

The multiple alignments have been conducted with MEGA6 (Molecular Evolutionary Genetic Analysis software) (Thompson *et al.* 1994).

Here reported the multiple alignments of the homolog protein sequences for *MRP1* (C1), *MRP2* (C2), *MRP3* (C3), *MRM1* (C4) and *MRM2* (C5).

Before producing the phylogenetic trees the alignments have been manually curated (data not shown).

C1) *MRP1*: multiple alignment of 9 amino acid sequences

```
>MRP1_Pseudo-nitzschia multistriata
MMTFNFSTVVLALVAAT--SFVSADYVCENQAFFKLDTKKKPSK----KSIDSLHTFMLD
SFQEAYKNNDDINMLSDKFESFSLGKDSSSILSA-LRGGNKDIDTLG----YGGGSY---
----YSQGRWGCNWCNV-DDD--AALGATIEAIDFGMALTTSSEHRLWEKLFCQKARTNK
DFKTISGCSIVLSDCHNENG------DEDAVEDETN----------------------
------------------------------------------------------ILSAVK
NMIN-----------------------------------------------
>CAMNT_0013081181Pseudo-nitzschia pungens, Strain cf. cingulata
-MMMKIFATALALVAAS--PVVSAIYHCESETTFQLEDSVKPSN----AAMNFLGTAMME
SFNEAYKNNPDIEMISDKSESLG-PASFGNIVTN-LRGKDNNKSNLG---GKMSTTY---
----VWGGNWGCNLCIV-DDD--AALGSFLSADDFGVALATSSEHKAWEKLFCERAHQMK
EFATMEKCAIVLKDCGEAND------YDGYEETFVD----------------------
------------------------------------------------------QVAKLIK
NPN-----------------------------------------------
>CAMNT_0008160915Pseudo-nitzschia australis, Strain 10249 10 AB
--MMKFATLALALVAAS-TPLVSAEYTCHGETYFQFEDGS-ISKP-SPEAIGYLDTAMVD
SFQEAYKSNNDIDMTAEKFDSLDLGLDFASIVNTALRG--GDNKNLG----RGGGRY---
----VTEYWWGCNLCVV-DDD-AFTLGTSIDGPEFGAALTSSMEHKVWEKLFCEYASAHE
EFDTMTSCSIVLSNCHKESASEHESESDVEAKSFVE----------------------
------------------------------------------------------QATKLFK
NSSN-----------------------------------------------
>CAMNT_0008148059Pseudo-nitzschia australis, Strain 10249 10 AB
-----------MIVSA-HPLK-------------------------QAIEYLGKAMVD
SFQEAYKNNNDIDMTSEKFDSFDVGPNFASIVNTVLRG--AGSTNLQ----RGGGRY---
----VTEYWWGCNLCVV-DDD-AVALGTSIDGPEFGMALTTSIEHKNWEKLFCEKASNLK
KFSSMTDCSIVLSNCHKESA------SDVKAKSFVK----------------------
------------------------------------------------------QVKTLVN
SSSN-----------------------------------------------
>CAMNT_0020483251Nitzschia punctata, Strain CCMP561
---MKLSFVFAVLSAVVAPVANANLYKCETSAVFDVEDDATKPQVPSKETLEWLAKELYD
TFHEAYAADSDVDMTSETFSKFTMKRETANEEDKTVSLSGKTVSLTDMISNLRGSRSSIY
GMVVYASSGISCRWCKL-DDD-AYALGKENGYTSIEDFVAASSEHKEWEKLFCAGIKKNP
EFASAKGCAIALTNCQDDGEANNDVDASVVGDNYLVSDMRAVQ----------------
------------------------------------------------EGEGVVYAVQ
TINQLLAGTDLETN-----------------------------------
>jgi|Fracy1|271829|estExt_fgenesh2_kg.C_300013
---MQFSTIALLLAALI--APSAAEYVCHSDASFNSDVDTMPSA----AAQKYLGSALVD
AFNEAYAGVDGVTMDYDDIEGFDTDPSV--SAAVLLRGGANLERRSR---SSRRRRS---
----RSTGGYGCNLCKTYDDD----ATAALSGIDFGVALLSSKEHVAWEKLFCAKGRANS
EFTSMTDCKIDLSNCHDDDE---DNVGGIPS-VVIP----------------------
------------------------------------------------SIVDILAIAS
MNTN-----------------------------------------------
>jgi|Fracy1|268858|estExt_fgenesh2_kg.C_50302
```

```
---MKFSTIALLLTTLI--APSAAEYVCHSDASFNSDVDSMPSP----AAQKYLGSALVD
AFNEAYAGVDGITMDSDDVEEFETEP----SAAVLLRGGTNLERR-------RRRRS---
----RSTGGWECNLCKTYDDD----ATASLSAIDFGIALSSSKEHLSWEKLFCAKGSANA
EFSSMTDCKIDLSNCHDDEV---DYVSAIPNKSGIPSQESLRRGLRKRRRRRRKRRTRSP
TNRPTKSPTNRPPRVTGKPLINGDDDDDYRPEPPLINRPKPFPDRTNRKQSFPELVGMTG
EEAKIYLKRYGKNGSNRKDGGICLFRNIDYNRVCWEEDENGLIIYAPFSG
>jgi|Psemu1|4292|gm1.4292_g
--MMKFCTFALALIAT--FTIVSADYQCESSTTFHWADGSKPSKP-SKADIDFLDKALLD
SFEEAYK-NSDLDMLSDEFESLEIGEDFASIVNR-LRG-NNDKPNLGEWLGSWYGTY---
----ISKMYIGCRLCEV-DDDMLDTTANSLGGSDLALALNTSAEHSSWEKLFCEKVHSRK
SFSTLTGCAIHLNNCETEPP----------------------------------------
--------------------SGEDDASHL-------------------------LKVK
NTIN----------------------------------------------
>jgi|Psemu1|291017|fgenesh1_pg.603_#_5
--MMKFLTTALALVAASPIASVSANYQCESDTSFVMADAVAPSK----EALEFLADAMKE
SFNEAYKRNPDVTMISDESEGFDSKPDLQTVVTN-LRG-NNDKPNLGEWVGSWYGTY---
----VAKMYIGCNLCEV-DDDMLDTTAAFLGGSDLALALNTAAEHSSWEKLFCEKVHKLE
EFASMTGCAIHLTNCETTTE----------------------------------------
--------------------DGADEEEHENF-------------------VEQAAKLVK
NLHN----------------------------------------------
```

C2) *MRP2*: multiple alignment of 12 amino acid sequences

```
>MRP2_Pseudo-nitzschia multistriata
------------------------MSLNNQANPCTVVAEVIDPN-------AVVTAPR
YPRARP----YSQRHLDGMEAQVVNQEQSQHVTFEADNKAAFFADSS---EGKNRFGKVS
NEKKGYNLLSSKWYWYAMIAMTLIFLVLVGYGFG----SGLFLLKR-----------GNA
TLNNDPAKDGS----TPNGSDPSSISDSSN--------TVQDRQDYKYSIVTLLGLPMVM
ERTSPQARAVEWLAYQDEPLFD---VTKET--SAEEQD----------------------
-----------------------RHHEILEQRYALVVWYFDQGGPTVWKTINRDESAGW
VEFGAGVHECDWKGVDCDYE---------------NGNTETGTVIGLRLSPALGLVLTGT
HLSTELGM----LTALRRVDFSDQRLQGTIPDEWASMTNLKSVILSKNQLQTTVPEWIGR
SSEEGGGWQNLEQLALDDNFLYGSLPSSMRNLRKLTHLELQVNPQLEGRFDELLFLEKED
-----EATGLPG----------QNLEILDLSNTRLKGKIP---------KITLPSIRSFR
--LWNLPGFSGTLPPDIGSWSNLEIFSIKEMPGLTGTLPTEFGSLEKLEILEVLDS-NFM
SGELPKELGNLSSNLKTINFRYTNQTGTLPVEWSSLVNLENLNLLQNKYLTGTIPSEYGY
MTSLR------------SLDLRGTSLSG--------------EVSQEVCAIDSMETLQA
DCSYKNDKSVGKIVCLCCLWCHDV-----
>CAMNT_0013082077 Pseudo-nitzschia pungens, Strain cf. cingulata
------------------------MPPENRDNFHAMVPEANGIRNHVKNVGVVKDAAV
TVAAYPGVKSYSGKDADCIEAQVVVERS---------TQVAFAENGGALFENSSNA-KDS
NKKRGFRSFSPDWFFCSMVAMTLIFLSLVGYGFG----SGLFLPSKAR--------IP--
TTNNDQVKDDA----ITAGDDAPSSVSVSS-------NSIQDRQDYKYSIVTLLGLPLIM
ERTSPQAQALEWLAYEDEPLFD---VSKES--SAEEED----------------------
-----------------------HYKAMLEQRYALVVWYFAQGGPTVWKTINREESAGW
IAFGAGVHECDWKGVDCDYD---------------DTETEAGIVVGLRLSPALGIVLTGT
SLSTELGM----LTALRRMDFSDQRLQGTIPDAWASMTNLESFILSKNQLQTTIPEWIGR
SPKNGGGWSNLDQLVLDGNSLYGSLPSSLSNLRQLTRLELQVNPQLEGRFDQILFSQEED
-----GTNRFLW----------ENLEFLDLSNTGLKGKLP---------RMTLPSIRSFR
--LWNMPRFWGTLPREIGTWSNLETFSIKEMPGLTGTLPTEFGSLEKLKTLEVLDS-NFM
SGELPRELGNLS-NLKTINFRYTNQNGTLPVEWSSLAKLERINLMQNRLLTGTIPTEYAQ
MTSLR------------FIDLRGTDLSG--------------EVSDEICAIESIEEFQA
DCSYQNDKNIGKIICVCCLWCHNS-----
>CAMNT_0008163717 Pseudo-nitzschia australis
------------------------MPSDAMVATAAADAAIGNENEDDAKMAAAATSPE
YPDAMPYYSSSNNLPELCMEAQVVAKQSKSFSHDINNSTACFGNNSGNNNNNDDDLGKDD
SNSRGKFSFSPNWFVYTMIAMTLIFLALVGYGFG----SGLFLPSKSA--------ISSG
STTNDPSKYNP----YASSSSSSSSSTSSNSNGDNVKNNRQDRQDYKYNIGTLLGLPMVM
ERTSPQARAVEWLAYQDQPLFD---VSVESSYTYEEEE----------------------
-----------------------IHKALLEQRYALVVWYYDQGGPTVWKTLNRDESAGW
```

IAFGAGMHECNWKGVDCDYKSVNIGVDG-------NSNNDRGTVVGLRLSPALGLVLTGT
HLSTELGL----LTALRRMDFSSQRLQGTIPDEWASMTNLETVILSKNQLQTRIPEWIGR
SSEAGGGWSTLQQLALDGNFLYGNLPSSLSNLRRLTHLELQVNPQLEGRFDETLFWQEQN
-----DGDDTNANANANANANTNNLEFLDLSNTGLKGKLP---------GITLPAIRFFR
--LWNMPGFWGTLPREIGTWSNLETFSIREMPGLTGTLPTEFGALQKLRTLEVLDS-NFM
SGELPTELGNLSSNLKTINFRYTNQTGTLPAEWSSLVNLETINLMQNGQLTGTIPTEYAY
MTSLR------------FLDLRGTDLSG--------------EVSQEICAIESIEEFQA
DCSYKNDKNVGKIMCACCVWCHNS-----
>CAMNT_0047397535 Pseudo-nitzschia fraudulenta

----------------MAGMTVVFVALVGYGFG----SGLFLPPETTPSDDRGFPIESG
ASVGASAAAGSSLLLAPPGSLPGVSSTSSS------LLSEFDRQDYRRTIATLLGLPVVM
ERASPQIRALDWLAYEDEPLIVSESVGGESRATSEETEG----------
----------LFRDLLEQRYALVVWYFDQGGPTVWKTLNREASSGW
IAFGAGVHECDWKGVDCGEGG------G-------GSDNTKRVVVGLRLSPALGVVLTGT
SLSTEIGL----LTNLRRMDFSDQRLQGKIPDEWAALSQLETVVLSKNQLQTTIPGWIG-
------EWTNLENLALDGNLLYGTIPSSLGSLQRLKHLELQTNPRLGGRFDTVLFQTEAG
-----GGGLRTS------------LDFLDLSNTDLAGELP---------PIELPSLRILR
--LWNTRGFSGTLPTQIGTWSSLETFSIKESPNLVGTIPTEFGLLEQLRTLEIEDS-NFM
SGRLPTELGNLSSNLEVISFRYTNQTGGLPVEWSGLVGLERINLMQNDRLEGTVPPEYSA
LTSLR------------FLDLRGTDLTG--------------EVPEEVCALESLEEFFA
DCSRKNDKTLGKIVCLCCVWCHGW-----
>CAMNT_0003599921 Pseudo-nitzschia delicatissima
----------MFRSNRKATTIREDQE----------YLDEEE
YPDAVVDPGNSAVVTAGVVLS--------------ASSTTFRREQHPDRNDQAQFDDDF
SKTRG------TCLDWFLLGMTIVLLTLVGYGFG----SGLYLPHRTVG--------DTN
ANSNDDATLGT---INGALESPGLFPLETN----------SQDYKYSIMTLLGLPIVM
ERTSAQARAIDWLAFDDEPLFD-----PNSMGDDHTQD----------
----------QRDRLAQRYALVVWYFDQGGPAMWTTLNREESAGW
IEHGAGVHECDWRGIDCDYL----------DGNDGIVAGLRLSPIMGLLLTGS
SVSSELGI----LTNLRRIDFSDQRLQGKIPNSWSLLTNLELVVLSQNQLQSTIPEWIG-
------GWTNLKHLALDRNQLYGTIPSSLATLQKLEQLELQENPQLRGRFDVLFSLDDDT
-----AGLQNT----------LEHLDLSNTDLVGALP---------NTTFPSLKFLR
--LWNTNGFGGTIPTEIGSWSNLEYFSLKENPQLVGSIPTEFGLLQNLTTLELLES-NFM
SGTLPTELGNLSSNLKIINFRYTNQTGSLPTEWSNLGSLQLLDVSTN-KLDGTVPPEYSQ
LTQLR------------FLDLRGTKLTG--------------EVPGGVCSIDSREEFLA
DCSTSKDVGVSKIECSCCGWCSSGNGGDN
>CAMNT_0041254931 Fragilariopsis kerguelensis
MSPSPSYPNYPDD----------NNLRHNAIVTTATASPSPGAFQGGIIDEEEGHFM
EARVAKKDHYEKKNGKYDFDDDNTNKDDSYNERYFNDNDAGDCDHKNTKKKTNNKDSSSS
LTFFSLFRDEGAWFMYTMVVIVVLFLALIGYGFG----SGLFIPDSRP----------AG
AADDT-FHINN----NTGG--ANSASLSSP------VSSKLARQDYKYQIMTLLGLPNVM
ERTSPQAQAIEWLAFEDEPLFV---VEDSATTTTTTTATTTNKTTVTDSSSSSSSNNSN
NNSNNNSNNNSNNNNNGYDDDYYEPYVGRLEQRYSLVVWYFDQGGPKLWTTLNREPSSGW
INFGSGVHECQWKGIDCDYDNVVIIGDN-------DNNDDNGIVVGLRLSPALGIVLTGT
SLSTELGL----LTSLRRLDVSDQRLQGTIPDEWSKMTELESLILSKNQLQTTIPEWIG-
-----NEWTKLEILALDGNLCYGDIPSSLSTLTGLRHFDVQQNQQLSGRFDKTLMMSSSL
MSSASASASAT----------ETIEYIDISYTNLTGSLPNSTN-------ILPNLRFFR
--AWNTRGLEGSIPSEISSWSNLEVLSIDDGPNLLGTLPTQLGLLTNLKTLEIQRS-FNV
FGTLPTELGLLT-NLEVLNFGASNHNGTLPIEYSQLTKLEKMDFMTNQALTGTLPIEYSS
LVNLK----------YLDLRATGLEG--------------EVSPEICAID-FELFNA
DCGSANDKSSGKMICACCTWCYAIAF---
>CAMNT_0020546941 Nitzschia punctata
----------MMNPSPPPEKTIDAHCTNEDIVFMPGAVS
VPFSGDGIPNDEFLGKKVVGHDEDDMEAKIAKKIMEEDDINHEYHRSVMKPDPNQNKRQR
----------DWYFVWMGLLTLIFVVLVFFGFW----SGLFLPNAAQ----------
----------EDVGSASS----------VERQAYMEELLSFFDFP-LL
EPGSPQEQAIEWLAYRDEPLPVP---SKEGGNSNNQYQ----------
----------RVRLEQRYALTTFYFAQGGPKLWSSINRHTWAGW
INYGMGVHECEWHGIDCEADN---DDGN-------NNDNERRHVVALRLNPSTGVVLTGE
SLSTELGL----LTSLRRLDFSSQRLEGSIPDEWKALTNLEMLALSKNQIQATIPEWIG-

-----DSWKQLKTLALNGNLVEGTLPTSILSLTNLKHLELQFNQKLQGRFDDLMVAIPD-
----------LEHLDISSTDLEGTIPS--------EVIMSNLRVFQ
--AWNTQKLAGTIPSAMGQWQSLEIFSIDDIPDMKGTIPTEVGGLSNLQELKIIRV--PL
TGALPTELGNLS-NLHTMVLSFLELKSTLPTEYGNLSNLETLDLNINSGLTGTIPTDYGK
LVNLK----------YFDVSTTNLSG----------TVSTDVCDLK-LDFFRA
DCPNKNPEP-NDIICVCCTWCL-------

>CAMNT_0020483817 Nitzschia punctata

----------MGLLTLIFVVLVFFGFW----SGLFLPNAAQ----------
----------EDVGSASS----------VERQAYMEELLSFFDFP-LX
EPGSPQEQAIEWLAYRDEPLPVP---SKEGGNSNNQYQ----------
----------RVRLEQRYALTTFYFAQGGPKLWSSINRHTWAGW
INYGMGVHECEWHGIDCEADN---DDGN-------NNDNERRHVVALRLNPSTGVVLTGE
SLSTELGL----LTSLRRLDFSSQRLEGSIPDEWKALTNLEMLALSKNQIQATIPEWIG-
-----DSWKQLKTLALNGNLVEGTLPTSILSLTNLKHLELQFNQKLQGRFDDLMVAIPD-
----------LEHLDISSTDLEGTIPS--------EVIMSNLRVFQ
--AWNTQKLAGTIPSAMGQWQSLEIFSIDDIPDMKGTIPTEVGGLSNLQELKIIRV--PL
TGALPTELGNLS-NLHTMVLSFLELKSTLPTEYGNLSNLETLDLNINSGLTGTIPTDYGK
LVNLMSVRPTCPVQLVRMFATLNWIFSGRIAPTKTRNPTTSFVSVAHGVCEID-SLHESL
RGDYKNVIS-DNRKLRKEIYCN-------

>jgi|Fracy1|257266|fgenesh2_pg.89_#_34

MAPEDSFDDDDDNGDGGGGKDDDRRYNNNNNNGLIITTTSSTPPGAFQGG--GGIEDRME
KAQAAKKDHDNSRNKNDMFDFDYDDDDDNNNH---NKNDS----HKSYNEVYFDGNCES-
----------AWFVYLMVVMTVLFLTLMGYGFG----SGLFIPNSGK----------YH
SVDSNGVIING----DNGGSALPNSGNSDR------YSYIVARQDYKYEIITLLGLPNVM
ERTSPQAQAIEWLAFEDEPLFV---VTKESTTTNETTT----------
----------NNNNDIELYNYESHLERLEQRYSLVVWYFDQGGPTLWTTINREPSSGW
INFGADIHECQWKGIDCEYNSNKYEGDD----------SEGGIVVGVRLSPTLGVVLTGT
SLSTELGL----LTSIRRLDFSDQRLQGTIPKEWSKMTDLESVLLSNNQLQTTIPEWVG-
-----KEWTKLETLVLDGNLCYGDIPSSLSTLTGLRYFDVQQNKQLTGRFDEILMMPS--
------SSTAS----------QAIEYIDISYTNLTGSLPSSNSNSSTTTLTLPNLRVFR
--AWNTRGLDGGVPSDISSWSNLEVLTIDESPNLQGTIPSQLGLLSNLKTLEILRSGTNF
YGTFPSELGLLS-NLETFNFRGNNHNGTLPVEYSQLTKLQRMDVQSNKVLTGTLPAEYGN
LINLK----------YLDLRGTSMVG----------EVSEEVCAID-FEFIMA
DC---NGRTSNEMICVCCTRCNKT-----

>jgi|Psemu1|98760|gw1.572.7.1 MANUALLY CURATED_stop codon added

----

----------MSN---LSKLT--RLELQ----------
----------VNPQLEGRFDQIFFYEDD----------EE----
----------DEDED----------EDEDED----------

----------TDED----------TEEDT
-----DSRRFPG----------ENLELLDLSNTGLKGKIP---------RMTLPSINSFR
--MWNMPNFWGTLPREIGTWSNLETFSVKEMPGLTGTLPTQFGNLEKLKTLEVLDS-NFM
SGALPTELGKLS-LLKTINFRYANQTGTLPTEWSGLVHLETINLMQNNLLTGTIPTEYAQ
MTSLR----------SIDLRGTDLSG----------EVSQEICAIESIREFQA
DCSYKNDKNVGKIVCVCCVWCHNS-----

>CAMNT_0000511607 Cylindrotheca closterium

-MEMASEQQPANSTEVTIVTSTX----------
----------PSSRAKGLVWXGSLILVGLCIAII----CIVLLRSSEX----------
----------EENNNXGVTIIVDR----------QGYYTYLLNTYSPLP---
--NTPQDQAIEWLAFQDEPLSGD----------
----------ELLSRLDQRYALVVLYYAHGGTSTWNSINDSSGSGW
INSGAGVHECEWKGVDCNNNN----------INAARQITGLRLSAEQGILLTGS
QLSTEIGCY---LTQLQSLYMADQRLQGSIPSDWKTLTNXXILDLSNNEIRSTIPDFFG-
------EFDDLQALLLGGNLLSGSIPATLADSN-IERLELHFNLGIXGRIEELLTSMPK-
----------LTYLDVSSTSFSGVLP---------STQVQKLQELH

```
--ALGS-DVSGTVPSEISTWSSLVNLNIG-HSHITGTIPSEFGLLPNLEGINLSYX--SM
EGTLPTELSRLD-SXSSLELRSSSFVGTLPTEYGQLKDLIVFDLALNGEINGGVPREYGN
MESLK------------ILYLHGTSLSG---------------TIDPAMCLLH-MIHFTA
DC---LERRAVELDCNCCTECYNLP----
```

>CAMNT_0050271705 Thalassiosira weissflogii

```
----------------MGRYSHRNSRTSHSARSQSTSLSSRPRDLYSRNHYDDDDDEDGV
YSHSQSRQGRQSRQSRNRYDDDDVEEASQAISTASSSVIDPNMPQWWNDRKKAERRAQIE
AERNAARLALETAKNGGAGAIPSSGASLSGRAFRGKGFAGAFIKESFTSRNSKNAALQKK
KQNQLDNLHNVDYMDVHRGSRANKKITADEELQG--RSLKDLVDRGTLMVGCCLILVVVI
AVTIPVALISEDEPYVAPPEPDPPTASPTSARLPDYWPIFDRLAPIAGGPEVLADPTTAQ
N--------RALNWIVYEDGMELGHSKDHLHQRFVLMVIYFISG--------PWTPVEGR
LEWGSPVHECEWEGIFCKDVNDLEEELQGRVDELLEVGREDGIKIDVPQQIVNRLELRQR
LVSGEVPAEFSLLYYLQHLDLENNQLVGALPTPLYKLFNLQTLFLEQNQLTNVDAIG---
------EYRHLENLALSKNAFQGSLPESFRNLNKLKTLYLHTNAWTGQVFEILKDFKD--
------------------------LELLDIAFNQFEGTFPP----------ELGDMKNLT
RFFAGHNRFEGEIPQQLSKCSNLKEFQVDGSHDVTGKIPTFFGRLEKLTFLKLDTC--AF
TGVLPSEIGNLK-NLTFLDVSSNYLEGNIPTELGGLESLVTLGLANN-DFEGGVPSEFGN
LKKLEK------------LYLTNTDLSG---------------AMPQEVCDLR-NRGNDT
VLELLQVPCDVECDTECCTMCS-------
```

C3) *MRP3*: multiple alignment of 6 amino acid sequences

>MRP3_Pseudo-nitzschia multistriata

```
------MNDESNKEWTYYFGRGGACNQQRKQSKVVGLLQKLALRYNQ--CHREEKRLFAK
CEVYNIVLEKGGSFFEITNKIAVDVTADEMKSTTKIMQVFRD----INKRNK-KSKGSAK
AKKKPSFTS-----------RTKKNKPSFCTLQTTRARESRTTKMKIKTQPKRRCVKPSI
IEAPIMNCVRDALLMASAKSQSLPDTRTNKXGLGTPVRGNGSE--------LKRLIAESN
HRFSTNNETFDLMAVEPPQLENSFSALIMADEVTT------------------------
---ELPTDSQMSVNLHLHKRVQRLENLVGMLMQQM----------------
```

>CAMNT_0008132293 Pseudo-nitzschia australis, Strain 10249 10 AB

```
------------------------------------------------------------
--------------------MPVDVTADEFESTTKIMQAFRD----IKKQCKLASTQSHL
GSSHLTTSSRKNKIKRMKTSSSQKNKTKRPRSE-TPSTKNPSPTMPSSP-PKPRCVKPSI
IEPPFIDCVRDARRMVDVNESIPEISAVSS--EMKP-SSNHKIPVSGGGSAQKPLVKKDT
IRNLSTDDMFDTMAVEEPKLESSFSALIMADEITA------------------------
---DLPTDSQKSINQNLHERVQRLENLVGMLMQQQNVD-----LERLL---
```

>CAMNT_0013109375 Pseudo-nitzschia pungens, Strain cf. cingulata

```
------MTDTAREQWVFYFGRGGIGNKKRKCSKVSELLQKLAPRYST--CHPTERRLFAK
NEVYNTVLNNGGAFFQIKDKLPVNVTANEHHSTTKIMQAFRD----INKNGKNVPTHSHV
G----THSP----------SPPRKTKTKAPKRK-VQSTKGPSPTMRLSP-AKRRCVKPSI
IEPPILDCVRKALRMFNADESIPDVGSDRA--ERHCENSNQASDVTCG---QTPLNK-DS
TRHFLSNDTFETMIVEPPKLENSFSALIMADEITA------------------------
---DLPTDSQKSVNQNLHKRVQRLENLVSMLMQQQNED-----LERLL---
```

>CAMNT_0003586753 Pseudo-nitzschia delicatissima, Strain B596

```
------MSETSKSEWDFYFGRGGGSNKKRQVSKANILVHDLAFRYSY--CSQTEKRLFAK
NEVYDTVVKNGGTFFLVANKEIIDVTKDFDDTISRIMQSFRD----INKSRRATSQTLST
NVEAKKVTN-----------NKSMQSTSLNTKRVPSLKKSSR-------PKRRCIRPSV
IEPPEIDCVRDTLRVIDTEKITPYTTTQLHIGELNILCDDGES--------RNTSINENG
LCQLAADDELEKMVVEKLKLDNFFSQVVMTDEIKARHPQQSLNLHLRVQRLENLVAMLMQ
IKSRLPTDAQRSANLNLHSRVQRLENLVSMVMQRKNDEKELKSAARIISNF
```

>CAMNT_0041255159 Fragilariopsis kerguelensis, Strain L2-C3

```
MMMMNNNSLNNNDEWNFYFGRGAENNKARKTSRANNLIQELAPLYIDPRTRPSDKQTFAK
QKVYDVVVNNGGRFI----EKDKNITADKVACLKKIMQGLRDCNKVVNKQCKPLPPSSPK
SLLKKTRDE----------SRVKKTPSNCLEIINVPSLKRTSSIIPVRC-QKRACVKLPI
VEAPIIECVRDALNVAKISGENQADTVNID--EKKAPSGIR-------------VAREE
KCDNENKDILEIMALEPPQLENSFSALIMTDEITA------------------------
---DLPTGSLEEYKQSVHARVQRLENLVAMLMQQKNEE-----FERLL---
```

>jgi|Fracy1|272356|estExt_fgenesh2_kg.C_440048

```
------------------------------------------------------------
------------------------------------------------------------
```

```
------------------------------------------------------------
--------MTSAL---KVDGEN-------------SLSGGG--------------GARQE
LYGSEDNDVLERMAAEPPKLENSFSALIMADEITA-----------------------
---DLPTDSQKNINQSLHARVQRLENLVAMLMQQKNED-----LERLL---
```

C4) *MRM1*: multiple alignment of 8 amino acid sequences

```
>MRM1_Pseudo-nitzschia multistriata
------------------------------------------------------------
-----------------------------------------------MNTPCTPPTK
YNKKAAVCISPEQMQES-EEVLLYRSNDGKSQLRTCRKRAATVLTSDAS----------
----------------TDGSGETFVTDDDGDSVSSGEAASSRKKNRTTRTVQRRK----
--------ATTTENPSPMPTKLTQAQIHSKSYTFPYKLFDLMEH-ATGEGGSAG------
------------SVVSWLAGGTTFCVHDHARFAAEFLPTHFGHHNFRSFDRQLNFWGFEV
LSPRHINNKSFGGKTWQHPFFQKDRRHLLGLVVRKTVSKSSPGGAKQNKRKPKGKPRQKQ
TVARTVSPVHTAPAGGRPGQAWD-------------------------------------
---------------------------ACPVG----------------------------
---------YDPRHHLLPLCPIEWLLEGDD-----GDDSVADGGAGVLPRTDLPVP----
---------SGSCSDDLD------PAVALDLNTSLNTVTMDDSDMEEAD-----------
----------LFLSIFEDHAEKKHNP-----HGSGGAEESD--------------ELFSF
VFGSS------------VDTDTILHAMNNSVSV-------------------
>CAMNT_0013145689 Pseudo-nitzschia pungens, Strain cf. pungens
--MSESFIASSSSSSCNTIVELLLFVANFFTCGIKKRFSPIAHIFLDTVSTNNTTAHCCI
GDCIYLQLILIFFPQLPSTQLLLLLLPLPLFLPPTPFSVSFSLLVHTKQTQHTVHTIAMS
TSSTSRACACTEQQQESSNEGLHPRETGGRRPLRKCRKRGIAAFAFEPQGPAKKKTKTNA
TTRSLALKSTLSFHIPEDDSTEQPFVTDDDEDSISSAETNRPANPRSTDKGVPPTAPTFP
PEEVCSPPRASPPPPTPLPTRLTQAQIHSKSYTFPHKLFDLMEDNAASDGNGNG------
----NGNGNGHTSIVSWSADGTTFCVHNHARFAAEFLPTYFGHHQFRSFDRQLNFWGFEV
ISPRNINNKSFGGKAWKHPFFQRGKRHLLQRVTRKLVIKPQQQQQQQQQQQQQQQQQXXXX
XXXXXXXXXXQQQQQQQQQQQQQQQKSLSSKKRMRGTASRTFSRRRRCLPPPPPPPLAT
AGTVPRHKYRMASTFHNPTSGWTLPRMISPVLGMGTDDGRGAAANADAAAATATAPSTLA
ANHSLTSGTIALAMDVLAPADPATFHENTR-----LAEAEAEAEATTETATTETETETTE
TETDLDPTPHGAAPPDLSE--EFLPLFPIEWFQQESGAPDTERGDNQAETTELLDRPNND
PEPGSVDCSDPALPLPEVGAASTAAE-----EGNPEVEEDSGDGNDDGDWGFVGKTFHDV
DLGEADLYLNIIEDSNGIGDQRATDGTGDCECECASLLCDCVLSSILLASV-
>CAMNT_0042638377 Pseudo-nitzschia heimii, Strain UNC1101
---------------------MVICIASQCKYELKSKSTDCNVPPMNTLIQDGSQRNRES
-------------------------------------TQIPSAVVKTAEKNECLRDFS
HLLPIIAARTQSTRHPSSGTGRGKKDNNSRTKKRKRVRGGTGTSTTKSS----------
----------------SSKIEQSVVSDDDCLGADLSVKSNRTKTRKRVKHTKTIT----
-------ISIPPKPPLPLPLKLKTSEIHSKKYPFPNKLYDLLSK-ASIDKRSSK------
-------------VVSWSSNGMAFVIHYHARFAAEFLPTYFGHTQLRSFDRQLNYWGFEL
VSPRNINNTSFGGKSWKHPFFQKDRRDLLKRVARKINGGSSSSQKCPSSRTNKTKKKTGS
SAKPSVKIISRDLNFHPKRGVQHVDGNDE---------------------------TAV
SATTNTSLALAKTFKNPSTVFELRPRIVSPVYGVSRSLN---------------------
--------SISPAQDLVLSSLDLVEWNGNS-----RKGNVARKNLSTLLAHNEVDFLPVF
SPTLFDTDNSSEQETGLMESPDNQETDPISSVEVFDSTDANSRDGEGGD--------AHF
PPKVGVFEGNRFHDIDLATGMDINVD-----MDMNMELNIDMDMDMDMDVDANEAELFLN
IIEDK--------NSNNVDEDLFSSHFDSQSAFDSAINTVDHFSVENVTCI-
>CAMNT_0011396591 Fragilariopsis kerguelensis, Strain L26-C5
---------MSSNIVGGDDDGNAVVVDGPSPTVIPARVVSSCMSKESDNSDYDSDNDNDS
-------------------------------------NSDRAINSNDAASLIPNTIVW
KGIKRMNGTKPSEKRKVTPPGLSSSASGSELSSSASASSSMAALVLAASTLS--------
-----------------DCANEKETTTEVENYHTKNRTNQQIGTQINPKTNICTRSVTSN
QLIKTAIKPKTNLFNQSTWNDNRGVMVHSKKYTFRDKLFDLLLDATSSGN----------
-----------TEVISWSSNGSVFVIHDHARFASEFLPTYFGHNKMRSLDRQLHYWSFET
VSPTSINNRSFGGKSWKHPFFQRDRRDLLELITRKKKKSKDPPSRRKHSSIDTHTDDERK
ASFEDHQKQLMNNTNNEQ-------------------------------------------
------------------------------------------------------------
---------ISSGREIDDQTSRSNSPSLMS-----HPRGPISDKVESCCSDDENDIII--
```

```
------------------------------------------------------------
------------------------------------------------------------
-----------------------------------------------------
>CAMNT_0000335841 Nitzschia sp.
MQGGGSIVAAARRNSSPSSTDNLKQLANVVMGQIHGSNGKDAASLVHGNEDGAPASGGMA
G--------------YAEAEMSRISPNSSNSAVASLGVTTPAASPTPSAGSKQGSSNGMD
LLMSVASAASADYALPSSERRKTKKRADGKSTPKKAKASPLATSGKKAVIASSESSASAI
DAQYAHKPDGYEHLFGQQGFASSSIPSGVSNSSVASPTASSDTPPPPPPKPTKSSNGGGN
----ANRRPSDAEAKRLRRNNKPHTDVRSETYPFPYKLFELIENAPP-------------
------------EICTWTKGGRAFIVHSHEEFRQKLLPMYFGHNVMRSFVRQLSYWAFDK
LSDPRVTLMSAGGCMWRNTYFQQGRHDLLKHVERVRVAGKRKKPPTSAPPPSSTTTGGNI
AASGNIQKKPKLTTTQTGPKSTKPK-----------------------------------
------------------------------------------------------------
---------PSKLHKQQQQVPMPVVPPAPS------VPTFPAQALEPQLAQDVVEV----
------------------------------------------------------------
------------------------------------------------------------
-----------------------------------------------------
>CAMNT_0042665851 Skeletonema marinoi, Strain UNC1201
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------MKTNEKKLQQQQQAFPFKLYEMLEYARASASGVGS------
-------------SISWLADGTGFVIHNKDAMMNDLTPMFFNQTKFRSFTRQLNLWGFLR
ADPR--------GENWKHKDFLRERPDLLKEIVRISVKSKSTMKVNMSMSSSGSSNIKAT
ASSQSAGSRRAAEVKGKRVNIETEKVVVA-------------------------------
------------------------------------------------------------
---------VAVADERVFPYSQEARARADS------NPTQFSYDYGGSAAVAAAASSAVV
ASPELRYLHPPQCQDSVFQYCSSNMKYDVQSHNQDSFSISSSVQLVSNS-----------
---------DEATTRTTTTTTQTQVASIQQNAVTNTTHATHRSSDVNATNQVQIQPFDDD
ELMYLAG-------IFEKDSPSQDDELRSILSLDRDCTIEDFMLAE------
>jgi|Psemu1|282196|fgenesh1_pg.4_#_2
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
-----------------MPTKLTQAQIHSKSYTFPYKLFDLMEQHATLDDHGQGHSHSHS
HSNSNSNSNSRTSIVSWSADGTTFCVHDHARFAAEFLPTHFGHHNFRSFDRQLNFWGFRV
ISPRNINNKSFGGKAWKHPFFRFDRRHLLEKVTRKLVTKNSKSNQKRQKQKQKQKQKLQR
DRCTSTSTSKDTEFTLLNNRSRDRDRNLN----------------------------AL
A--------------------TLPRTISPVRNVGVG-----------------------
---------VGVGIGVAPIPPPGTHHSSCR-----GLMASASATAKGTKNTDEYASYATT
---------DGMLSEDL-----FLPLFPIEWFPESEHEHEHEHERGGAR----------E
PQP------HPATLLPTATATASPTE-----TASPEPHEGT---------AHVHPPQVFF
DFGEP--------HEPRVGPESHARGSRGWDTG--------ALVGALGAAT-
>jgi|Fracy1|250502|fgenesh2_pg.31_#_101
-----------------------------MNVPTTPTTTTERAATQISLDETIK----
----------------------------------------LELISPFSTGLIPIRRRVL
RSSSFKNTNTCGSNNKISNDRRRKRKSAGATTTGSAVSTNAQCIVSDDDES---------
---------------ILNKKKRKKDVTINYNTSTNTSNPAPTAFSTKNVGGNATIS----
--------SVSSSSSSSPSSLNSQAQIHSKKYPFPHKLYDLMAKTDSRD-----------
-------------VISWSADGTEFVVHDHARFASILLPIYFGHNQMRSFDRQLNYWEFER
TNTNKISNKSFGGKSWKHPFFQKDRRDLLKEIVRKKIKNGPASTRRYTFNDGDDKKKNKK
KNKDKNTAATKIKKEQRGMEEVLESVIPG------------------------PTHANV
RKNMLLKIQSINRKGENLASMGLRRNIISPVVNQESSSLSLSSSSS--------------
-----SKESLPPLPELEIEIAPPIQLRRNTNLVVFQESSLLSTSSSSSLSSEEEESLQPL
SEYVRTPSEMIATALYEDIEDEFLPLLHIDCFNDVDVDGDGDHKGDKNEDKHNMHNSSNN
TDGGMVYQWDHNGIVFGGKSFHDVKSGVYDYSVPCLFEAAINDSDDDIGDSESEREKKRD
VLKKLTVPKLKDRLLEAGAKHSQFRSLRKSDLIELLIQMDQDLPTTTTTLVH
```

C5) *MRM2*: multiple alignment of 11 amino acid sequences

>MRM2_Pseudo-nitzschia multistriata

```
---------------MLATSDPYPSPHAKPEANVEYGNVQRFPVESDDGYRAQDQRKNR
NSYRPLFITLYVMLLLATAGLAFAVTRYVQTVKNKHKSPPSQDT--TSDGSTDVDSSSPA
TLAEVDPDRDLIAYRSDIEYILFTEMDEGLSTDFVEGPQKRAIDWLVSDDLVLNSTEVR-
------AMAEY---------------IKNGDEDSVSTVPLVQRYALMVLFFATNGELWSD
SSWREL----------VNVPECRFMGIECD-LEGHIN----------------------
---------------TLDVGYRKLRGRLPGEVGMLSMLTSLNVESNNLEGTIPSFLYNK
LTKLERVDMRNNGFLSTISSD-ISKLTNLKALYLGELF-LTGEVPTDAMKSLSSLEEISI
SHATEMTGPLLEFSEHWPNLTYFDILRST-FTGTIPTTIGTNTNLKYIWLEQTSMNTSIL
PTELGLLPNLKEFILDNLVVD-------DTTGTIPTELGNCQALTSLHVD--KFRGPIPT
ELGRLTN-LKFLSMTEGGLTGSVPSELGLLTNLNEMYLYNNRLESSLPSALGNVQG----
-----LKILDISMNNLTGS-IPEGICRSPS-IGIKRDCFIDKCDCCSLYCIEG-------
-----
```

>CAMNT_0013118197 Pseudo-nitzschia pungens, Strain cf. pungens

```
---------------MPLPSASSNPTIQADVVFMPDKAPSTEGQPIPVEMGRVDRQSAGR
NAYRSLFVTLYTMLLLATAGLAYAVTRYIQVKKSAHILRSSQGTGTATEELTDIDSSSPK
TLAEVDPNRDPASYRSDIEYILSREIDGDCSTNFLEGAQKRAIDWLVFDDFVLKSKAVRK
MVESLEAIAES---------------NAAESKDSIPTFPLLQRYALMVLFFETNGELWSE
QPWTEM----------VGTEECKFGGVECG-LEGHVE----------------------
---------------VLDLGFRKLRGRLPREVGMLSKLTSMNVMGNNLEGTIPSFIYHK
LTNLASLILSKNEFYSTISPD-IAKLTNLKALHLTELQ-LTGQLPADAMKSMSTLEHILF
ADSTKLSGPILEFSAHWPNLTVLDLYQSS-FTGTIPATIGTNTKLEAIWLRDTEMDTSVL
PTEFGLLSNLEQLFINAKVPE---------GGRIPTELGNCPQLRWLHLD--GYGGRIPT
ELGRLTR-MEGLSLSRGTVTGTIPSEIGLMTSLMSLYVTGNQLVGTLPSEIGNLR-----
----EMQDLRLNRNSFSGT-IPTGLCQG-PPKQIQRDCGVD-CECCSTPCSRLE------
-----
```

>CAMNT_0042654255 Pseudo-nitzschia heimii, Strain UNC1101

```
MASKDSNSSDNLQSETVLQAQVVFAPDDLNNQKDHALPNGTEDN-FYPIESDPEQDKGKV
VSDRPLFMTLYAILFLAITGLGFSVTRYVQVKRNVRSSQ-------SSQEIASDNNTLSA
TAEDTDLDRDPVAYRSDIEYILSSEIYQDSTISFLQGSQKKAIVWLVYEDRVLSTTKIRE
MVDYMKSNNTGNDD---------XXXNNTGNDDVVPTFPLVQRYAMMVLFFETNGELWSE
RSWAES----------TNLNECKFVGVECD-MEDRVV----------------------
---------------VLDLALRKLRGRLPEEVGLLTRLESASFLSNSLEGTIPSFIFNK
LTNLHSLDLSGNDFSSTISSD-ISKLTNLRRLHLNELS-LTGELPVDAMKSLSNLEQLAM
THATRMNGPLLEYSISWPNLTSIDIYQSH-FSGTIPTTVGVNTKLERFWLPGTKMDLSSI
PTEIGLLTNLLQFSLVSEMHRMD-------TGTLPTEIGNCRALQYLQMVGSNYHGTIPT
EFGRLKN-LTEIYLNRGMLTGSVPSEIGQLTNLQSLDISRNQLQGTLPIELGTLQD----
-----LKSFEVYRNNLSGT-IPSNICNS-PYISIIRDCAVDKCKCCHLPCIVNNNYNSKK
GQDRM
```

>CAMNT_0003619981 Pseudo-nitzschia arenysensis, Strain B593

```
------------------------------------------------------------
-----------------------------------MH-------VAQDGAA--VELLS
PVEQADPNRNPDVYRSHIEEILLQELHEECSTNFLEGPQKMAIDWLVYEDLVLNSTLVES
MA--------------------------NGD--EPTFPLVQKYALMVLFFGTSGELWSG
QSWNRM----------AEVSECKFMGIECD-LDDRIT----------------------
---------------IVDLGYRKLRGRLPEEVGMLTNLYSFSVISNSLEGTIPSFIYNR
LTDLNTLDLSKNNFASTISPE-ISMLSNLQVLHLMELVSLTGEIPVDAMRQLTSLEHLVI
TFAQNLRGPLLEASSSWTNLNTFDVYLSG-FTGTIPKTIGQSANLEYLWLEGVQMDASTI
PSEIGLLTNLKGFILD-SKSPLD-------GATIPTEIGALPNLEIFGGR--GFTGSIPT
EIAHLAN-LQELILEAGSLTGNLPSEIGVMTNLSHLRIVLNEIEGTLPSELGNLQK----
-----LELLEVYFNNLTGS-IPSGVCKN-PDISIDRDC-IQECECCGTKCRYKKDK----
-----
```

>CAMNT_0003575869 Pseudo-nitzschia delicatissima, Strain B596

```
--MEDLDAYTTNPGAILQAEVVVSNVRRNQNGREDERPKEGEELXXIETQMSLDEKDTTD
NSSSRLFLVLYTILLLSIVFLVVALTRYIRVKRSVVPIP-------VSQDAAS--AELLS
PVEAADPDRDPVTYRSHIEEILVRELHDECSTNFLEGPQKMAIDWLVYEDVVLNSTQIGN
MA--------------------------VGYDLEPAFPLVQRYALMVLFFATSGELWSG
APWNLM----------VDIPECRFMGIECD-MDDRVL----------------------
---------------LMDLGFRKLRGRLPEEVGMLINLRSFSVLSNSLEGTIPSFLYNR
```

LTNLDVLELSKNEFSSSLSSD-ISRLSNLKILHLAELS-LTGQIPVDAMKQLTSLEQIVI
TYATKMKGPLLELSENWPNLNVVDIYMSG-FDGSIPTTIGQNTNLEMLWLDAVPMDASTI
PTEIGLLKNLRGFVFRPSRNGVG-------GATIPTEIGECTNLDTVDWN--NLSGFIPT
EIGKLTK-LKEVIFEDGTISGSIPSEIGLLSKLTLLSILGNQLKGTLPSELGTLSN----
-----LEVLDVYANNLTGN-VPSGVCENNMDIFIDRDCSIEECECC-DKCRGERRY----

>CAMNT_0011214887 Fragilariopsis kerguelensis, Strain L26-C5

-------------------------MRNVGPGSSTTDPDG------ITSSSSSSSSSSN
TLEESDPNRDPIKYRSDIESILSTVVEEFT---FLEGAQKKALDWLVFEDLVLTSADVKA
MMGSTKNDDNDN------------GSGSDDDGVGVPIFPLVQRYALMVLFFETNGELWSD
TSWSDL----------IHVHECNFMGIDCDGNKGQAV----------------------
---------------NLDLQYRKLRGRLPDETGLLTQLTAVNLMANNLEGSIPSFMYNE
LTNLEFLDLSRNDFTSTISSD-ISNLTSLKSLVLNDLS-LTGSVP-ESLKSVSTLEDFSI
WHGGKMSGPILDFMSYWPNLVTMDIYQSS-FTGTIPTSIGINSNLKLFWIEDTPTDITDI
PTELGLLTDLEEFAIGSRIGLDG--------GTLPTELGSCTSLRNFQIDGNNYTGTIPT
EIGLLTN-LASLVLSGGRLTGTLPSEVGNLTNLVFLLLHENELKGILPSEMGSLTA----
-----LDTLDISDNDMTGTGIPESICSSHPDVWID--CKVEKCSCCYRPCFSGSG-----

>CAMNT_0008239669 Pseudo-nitzschia australis, Strain 10249 10 AB

---------------------------------MLSKLTSINLESNHLEGTIPSFVYNK
LTNLVFLSLSSNAFSSTISPD-ISKLTNLRGLHLAELF-LTGQLPSESMKSLSTLEAIVI
PSATEISGPLLEFSESWPNLTIFDIYQSD-FTGTIPTTIGTNTKLTTLWLEESSMNTTVL
PTELGLLSNLQEFASDNMITGRGTITSASTSTTIPTELGNCHAMRWMHLD--RYGGLIPT
ELGSLTN-MESLSMSDGTMTGSVPSEIGLLTKLRELELFNNKLKGKLPSSFENLQA----
-----LMDFNIFRNDLTGS-IPSGMCDTTLNIAIQRDCSIDQCKCCFLPCKSSAT-----

>CAMNT_0049115329 Staurosira complex sp., Strain CCMP2646

----------MTDTVNGSIQGVVDDISAATSHAVVVDEQRDNEEMAASATESSVHQDDTA
NAVVQNASSTNGEILQENSHGNGNRLVTAPVEWSSRRRLSNSTAGFNTATREQPNPSSGD
TQATSTSTSNESSVKNNNICCGMSRWKVVTITTTVLVLVIGIVIGLVFPVRHDATSNTTG
LPPDFSTTNQAILRVRTYLLQETNWSDPLSLMDPTSAQTKAVYQLALEGGQAPVQQQRYG
LLVAWYGLGGGDRSSGLGRQECDWNLVACNNEAQVTS----------------------
----------------LLFSKQGLDGTLLEEVALLSNLEYLDLEENTVKGQLPAALYS-
LTNLVTLSLGFNEFRSELSDG-IGSLTKLENLRLHDND-FSGTLP-DSIKSLTNLKSVEL
WRT-NLAGPIVQYAASWPNLEVLSIAENFNINGTISSEIGTLSKLRELILQRTHISG-TV
PSEVGLLSNLETFSVG-LYAIFG--DSVTVSGTLPTEIGNCQKLRKIVIEASNMTGPIPP
SIGRLTSTLKILSLSDSNFTGSIPDSLQSLTDIHEIYLANNAFSGTLPEWLGSFEE----
-----MRHLQVYWNQFTGS-VPTGLCRASLKFYYD---CVLECDCCG-PCGPLV------

>CAMNT_0004089223 Cyclophora tenuis, Strain ECT3854

----------MMTFAVIGAVIFVTTTPLGGGQSKELNEDGGV
ENPEELLAGMRGFLVQELSGDEATAFIDP------------------------------
-------TSTQSSALD----------WLVFEKQYEGN-----
----------------WWQGQRLDPSQVRQMDPELAQQFLQRYSLITLAFACGGEKWVR
IRLRWT--------ESAHLHECDWRGINCNEDDVVTD----------------------
----------------IDLIRVGMSGTIPQEIGLLTKLESLNLRKNRVAGLIPSSLYR-
LTTIXFLNLGENSLVVPMNDTNFDLMTNLEYLSLDDNT-ITGVLP-PAFGTLRRMKVINL
EFS-DLSGQILDSLLQLPLLEEIRLMYTR-IKGTIPTEIGSLSNLRYFRVGSHELSG-TV
PTELAKLSNLVSLIITGNVEYAG--QENKMTGTIPTELGLISPLVDLEISETEIGGTIPS
EIGRLAN-LGYFHLRGGNYTGTIPSELRNLTNMAAVYLAHNGLEGEIPTWFSTFPN----
-----LILLELAFNNFDGR-IPEELCRPSLRIAHS---CDQFCDCCYSACLPGVG-----

>CAMNT_0042604521 Asterionellopsis glacialis, Strain CCMP1581

-------------------------MEDIIQQVAEPEKEKDKQMTAAIHNAMKEASPD
AELKPGAYSVRAPAWSFFRGARESISMSRRAKFGTKRARSGLNNAVTDTGIEDNMGDFID
EELSWWEALNIGTICVGCCIIAIVLVGVLIPVLKLNGDEDEPINAGPYEKVLFPLSGKDS

FKDTTSPQSLAFD----------WIVYTDKLPIDSPNLIQRYSLMVMYFSNKGGDWIV
HQTQWG----------SSDHECNWDYIECSVCSKENLNVVGSTSQALGTGGRLRHRSLEN
VFALPGKDGTQCVSSLIFDKAYLKGELVLPDELSQLEYLYKFDVAENNLSGKIPDNLYS-
LSTLQHIYLERNQFTGMISSN-IGNLNKLKVFSVFQNT-LTGQLP-ETMKTLESLEILWT
SRNIGIGGPLSEFIASWPNLVEIDVSNCA-FQGSIPEEIESLSNLKGLYLYDNKLSGS-L
PTELGLLTTLSRFVISRNRSMTG---------SIPSELGALRNVEELRIDSIDLDGSIPT
ELGNLSR-AVIINLAENNLIGNLPSELGRCSSATELSLFNNQLEGTVPSELGKLSQ----
-----LATLYLQRNNLTGE-VPGQVCSLRDELLET---FITSCSDGDLEC----------

>jgi|Psemu1|202069|e_gw1.290.19.1

-----M----------VGVHECKFTGIECD-SEGRVT----------
----------VVDLGYRKLRGRLPDEVGMLSKLTAVNLMGNNLEGTIPSFVYNK
LTNLGSLMLSSNEFHSTISSD-ISKLTNLKKLYLGELF-LTGQFPVDAMKSLSSLEHLAV
SDATEVSGPLLEYSSHWPNLTYFDIYESM-FTGTIPTTIGTNSKLQ----------
----------LPTELGNCRSMRWMMFS--EYGGSMPT
ELGELSN-LQHLTMVNGMATGTIPSEIGRLTNLVYLSVFNNRFEGTLPTELGNLGGGGAS
NSNSKLQFLNLYKNNFSGT-IPSGLCDGGRPMQIDRDCGLA-CDCCSKLCNRAG------

APPENDIX D

CT study and REST analysis for the reference and target genes used in the 24 h time course experiment

The Appendix D reports the data about the CT study (section 1) and REST analysis (section 2) for the reference and target genes used in the 24 h time course experiment.

1) CT study for the reference and target genes used in the 24 h time course experiment

Appendix D presents detailed graphs for the CT study for the reference (Figs. D1, D2, D3) and target genes (Figs. D4, D5, D6, D7) used in the 24 h time course experiment, which show the CT values (mean of the technical triplicate) trend for each reference gene, considering the six time points and the six biological samples.

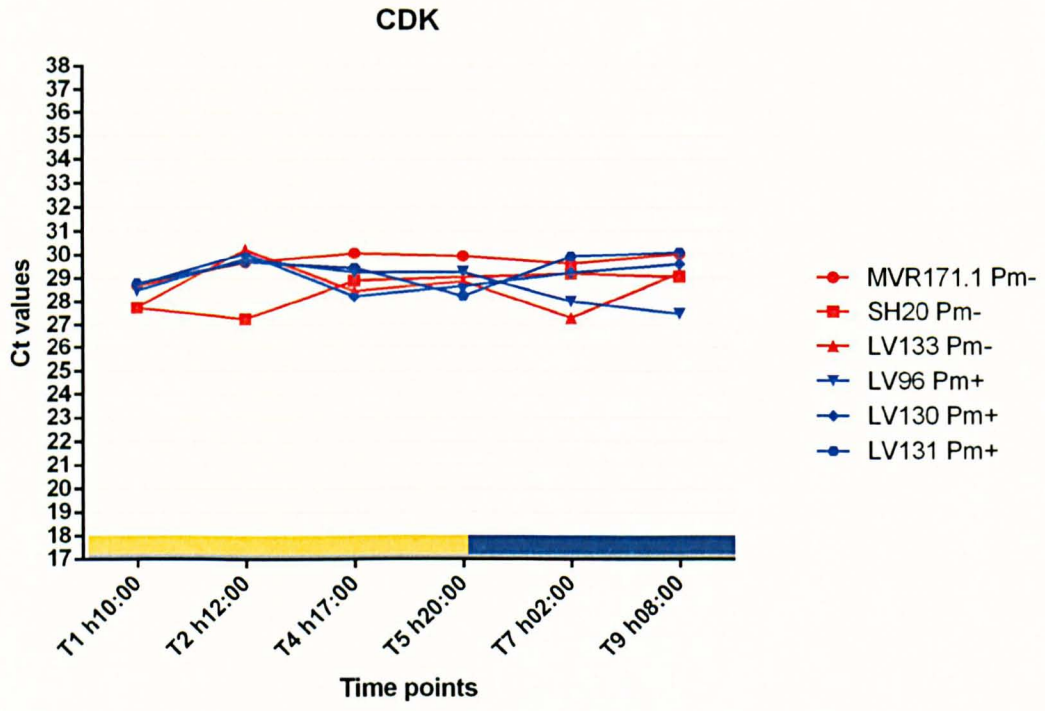


Figure D1: Expression levels of the reference gene CDK taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values) for each biological sample.

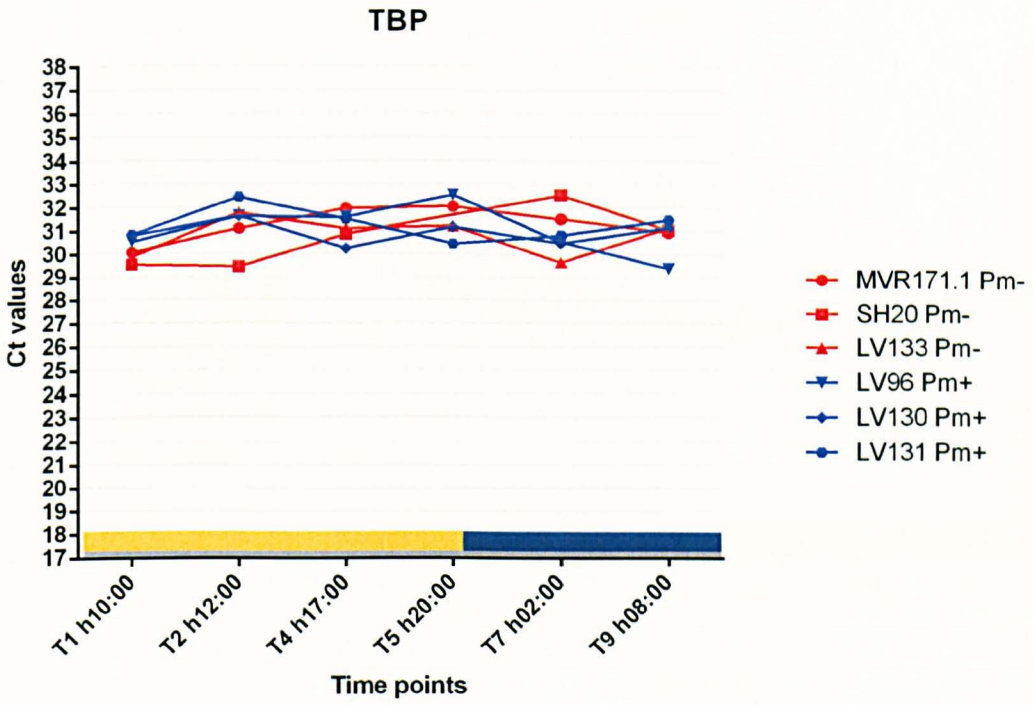


Figure D2: Expression levels of the reference gene TBP taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values) for each biological sample.

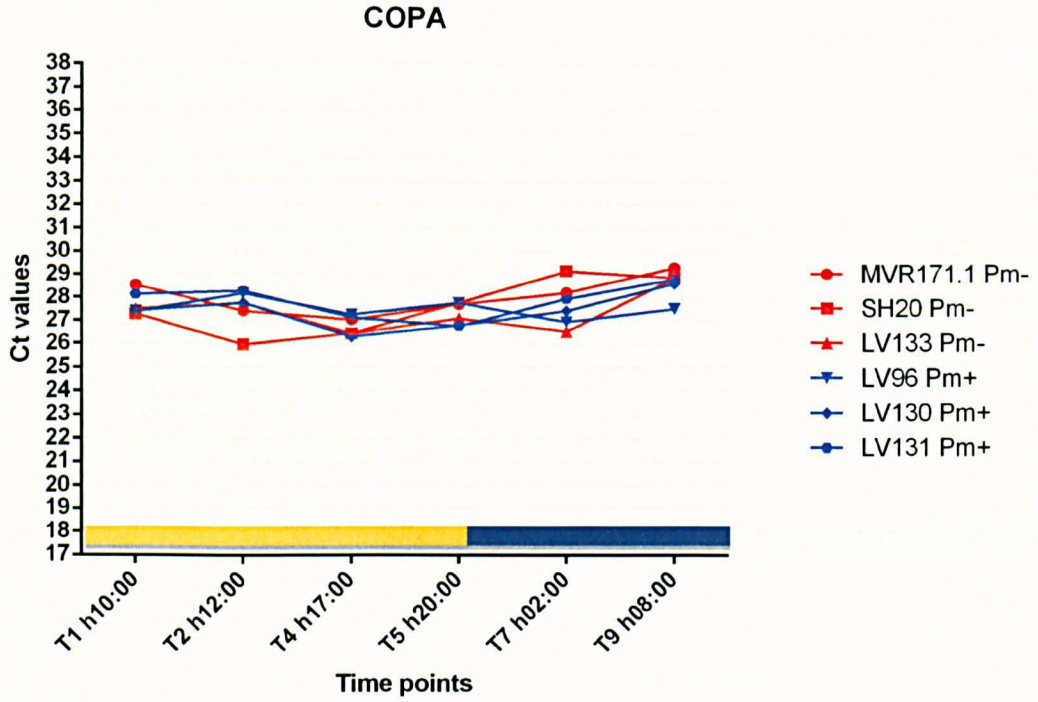


Figure D3: Expression levels of the reference gene COPA, taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values) for each biological sample.

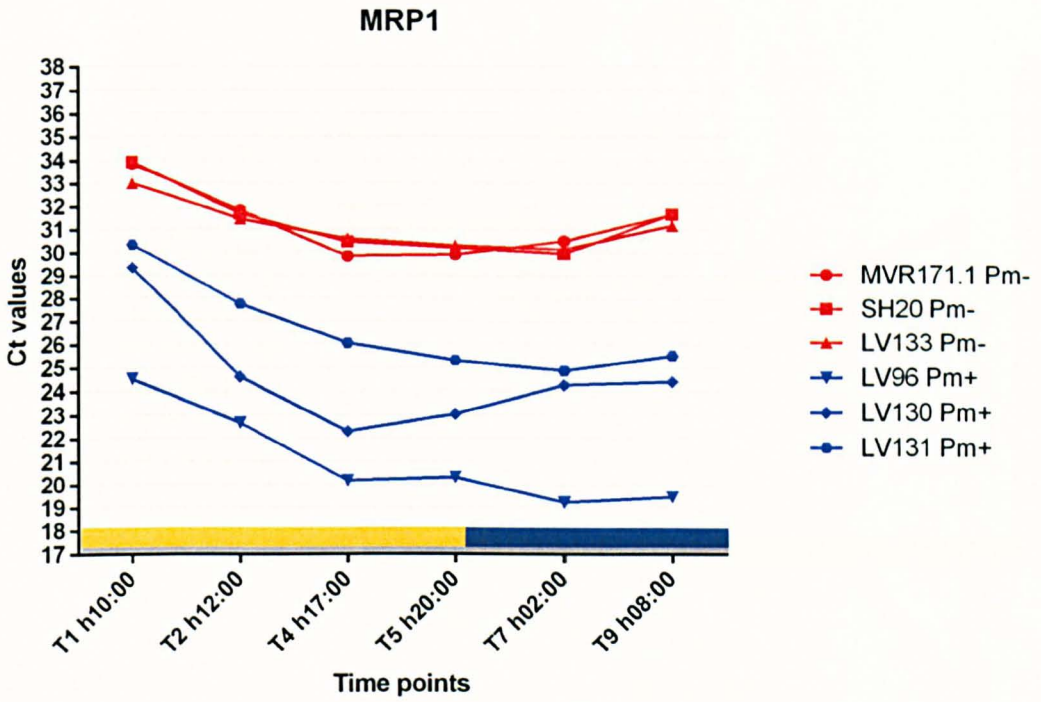


Figure D4: Expression levels of *MRP1* taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values). Symbols with connecting lines represent the expression trend of *MRP1* for each biological sample.

MRP1 presented high variability among the MT+ biological triplicates. In one/two samples (mainly LV96 Pm+ and LV130 Pm+), depending on the time point (mainly T1 h10:00), considerable differences in expression rates were observed (Fig. 3.11).

In the following graphs, for the remaining MT-biased genes, marked strain-specific variability was not observed for MT+ strains nor for MT- ones.

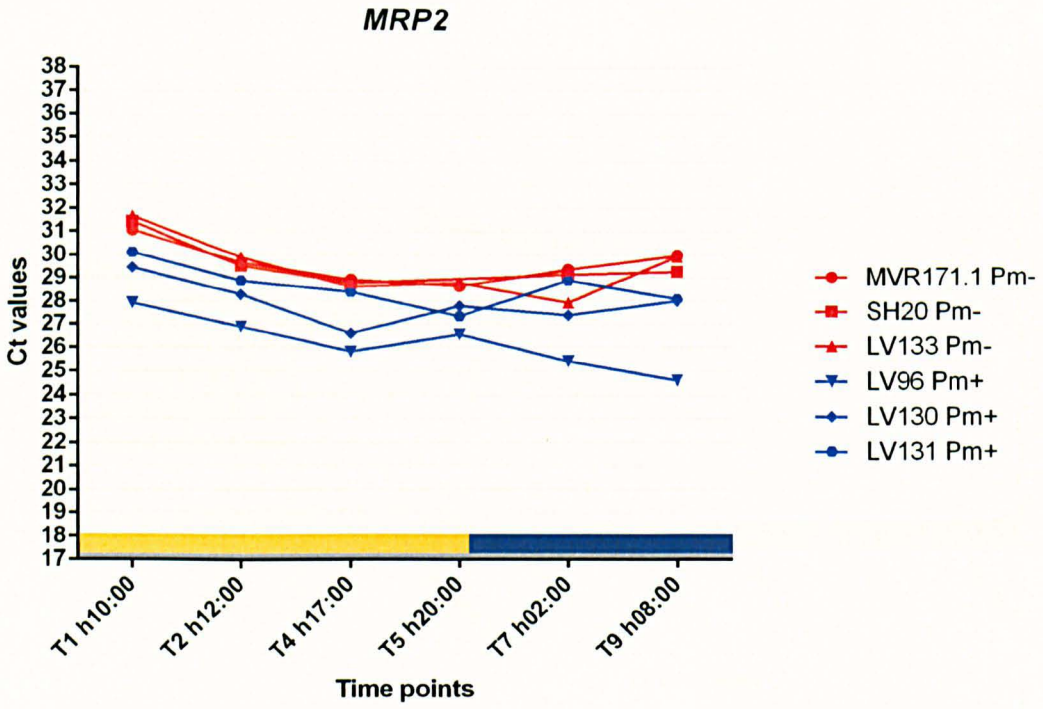


Figure D5: Expression levels of *MRP2* taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values). Symbols with connecting lines represent the expression trend of *MRP2* for each biological sample.

MRP2 displayed the same temporal trend of expression of *MRP1* presenting a decrease in expression for T1 h10:00, but less pronounced than *MRP1*.

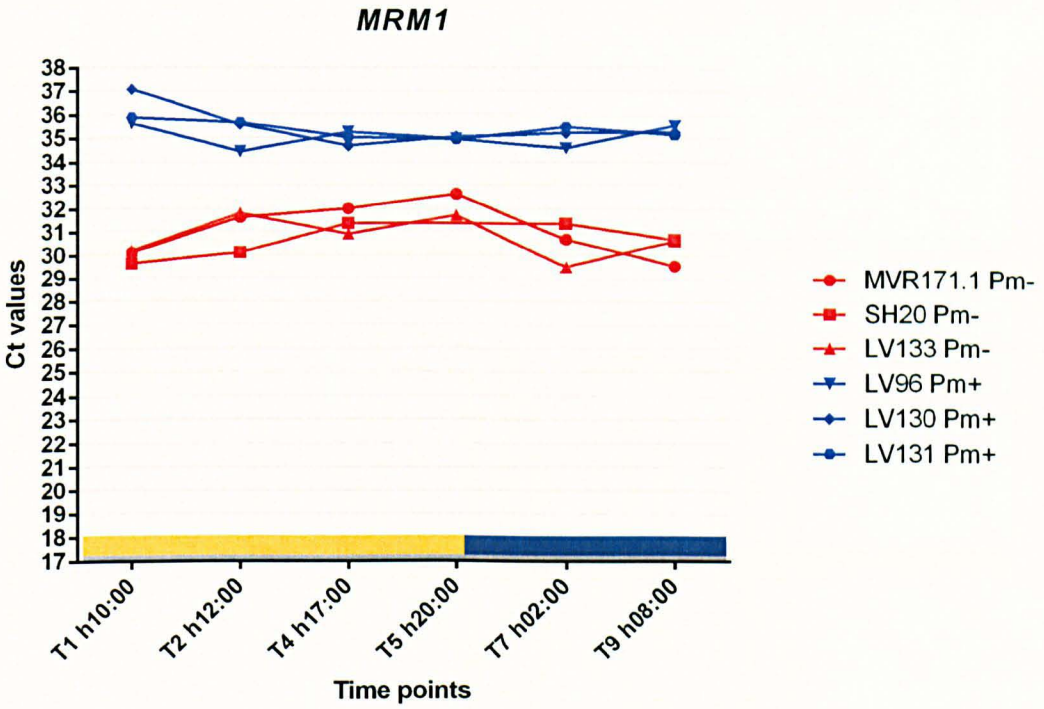


Figure D6: Expression levels of *MRM1* taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values). Symbols with connecting lines represent the expression trend of *MRM1* for each biological sample.

MRM1 resulted to have a general low expression rate (see CT values of 29-32 among MT-samples). Its expression trend looked quite uniform; however in T4 and T5 a decrease in the expression level was observed.

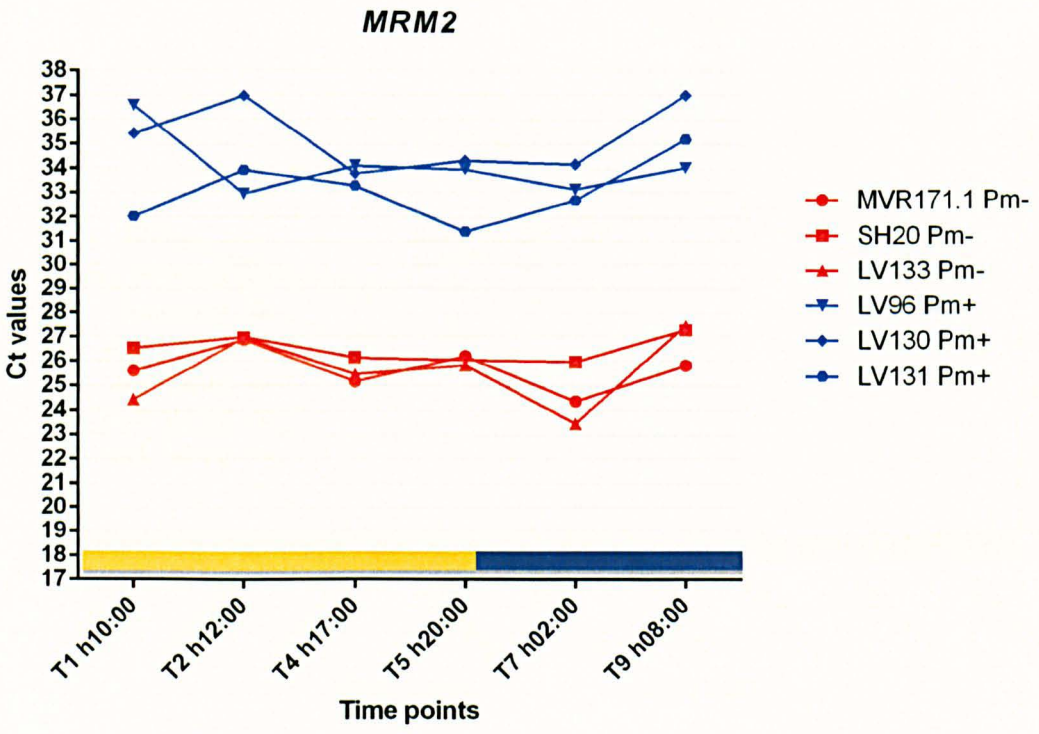


Figure D7: Expression levels of *MRM2* taking into account 6 time points during the 24 h cycle. Values are given as qRT-PCR cycle threshold (CT values). Symbols with connecting lines represent the expression trend of *MRM2* for each biological sample.

The CT values of *MRM2* only in T7 (mainly LV133 Pm- and MVR171.1 Pm-) were deviating from the uniform trend of the other five time points.

2) REST analyses for the target genes used in the 24 h time course experiment

The significance in expression variation was calculated setting the first time point (T1) as control against the other time points, set as conditions, and normalized over the expression variation of reference genes whose expression levels were not regulated in these specific experimental conditions. The relative expression ratio (R) of the targeted MT-biased genes was computed separately for each biological replicas carrying the same mating type (3 MT+ and 3 MT-). Since REST-MCS gives the possibility to test only 7 condition per time, one biological triplicate at a time was analyzed, so each sample had as reference condition

its own T1. The expression ratios obtained through single REST analyses were plotted together (Figs. D8, D9, D10, D11).

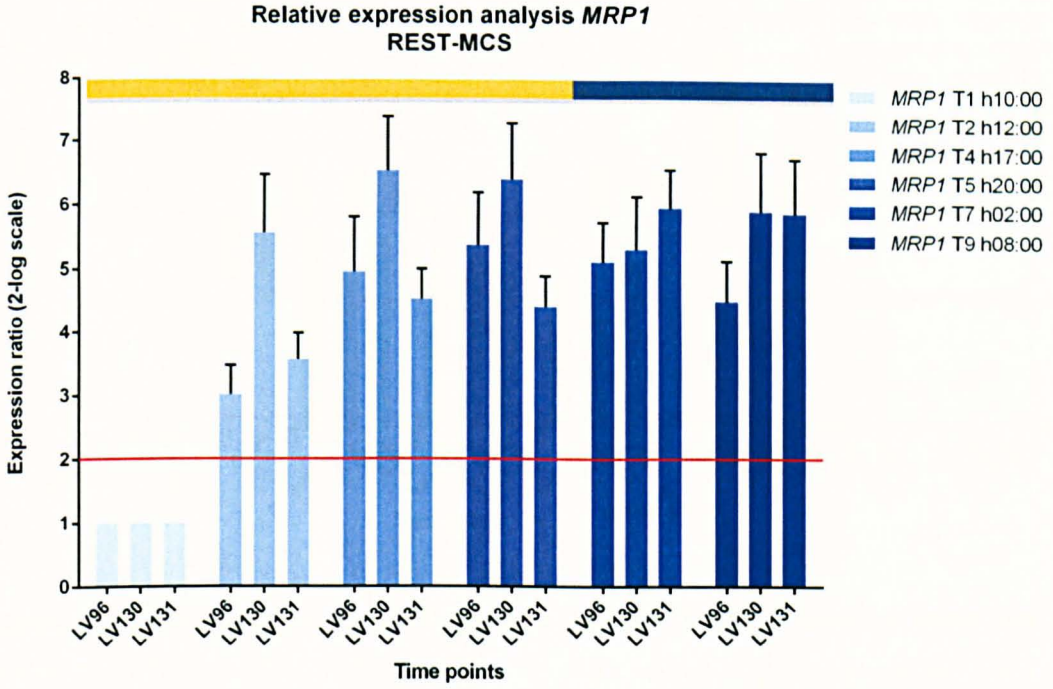


Figure D8: REST analysis of *MRP1* obtained by fixing T1 as reference condition; normalized against three reference genes *CDK, COPA, TBP*.

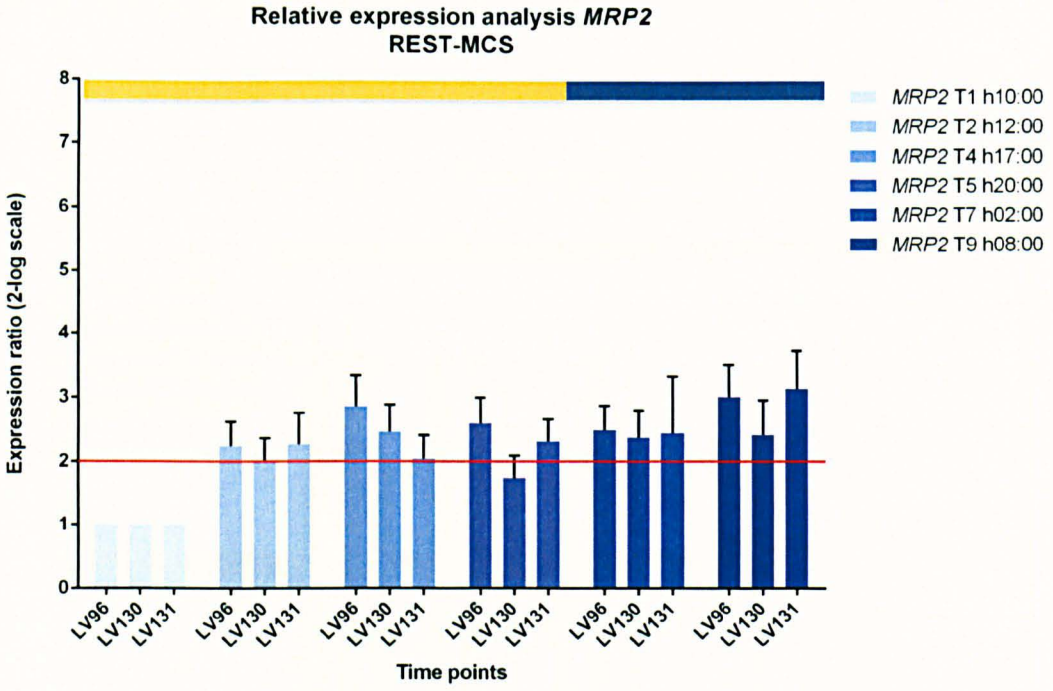


Figure D9: REST analysis of *MRP2* obtained by fixing T1 as reference condition; normalized against three reference genes *CDK, COPA, TBP.*

Expression rates of *MRP1* were significantly up-regulated in T2, T4, T5, T7 and T9 in respect to the control set as T1. The same behaviour was detected in *MRP2* but the expression ratio was lower. So, both *MRP1* and *MRP2* presented at the beginning of the experiment, 2h after re-illumination, low expression rates that after T1 tended to exponentially increase till T5 and remained constant until T9.

COPA was excluded as reference gene for the normalization of MT- biased genes because it resulted to be differentially expressed in some of the time points.

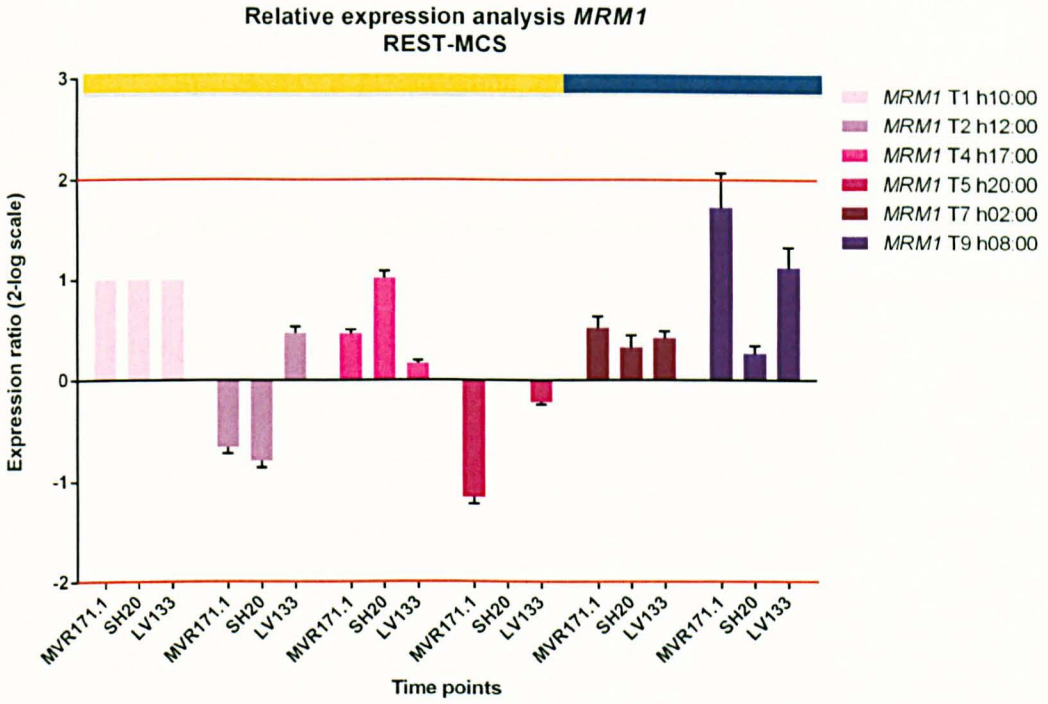


Figure D10: REST analysis of *MRM1* obtained by fixing T1 as reference condition; normalized against two reference genes *CDK* and *TBP*.

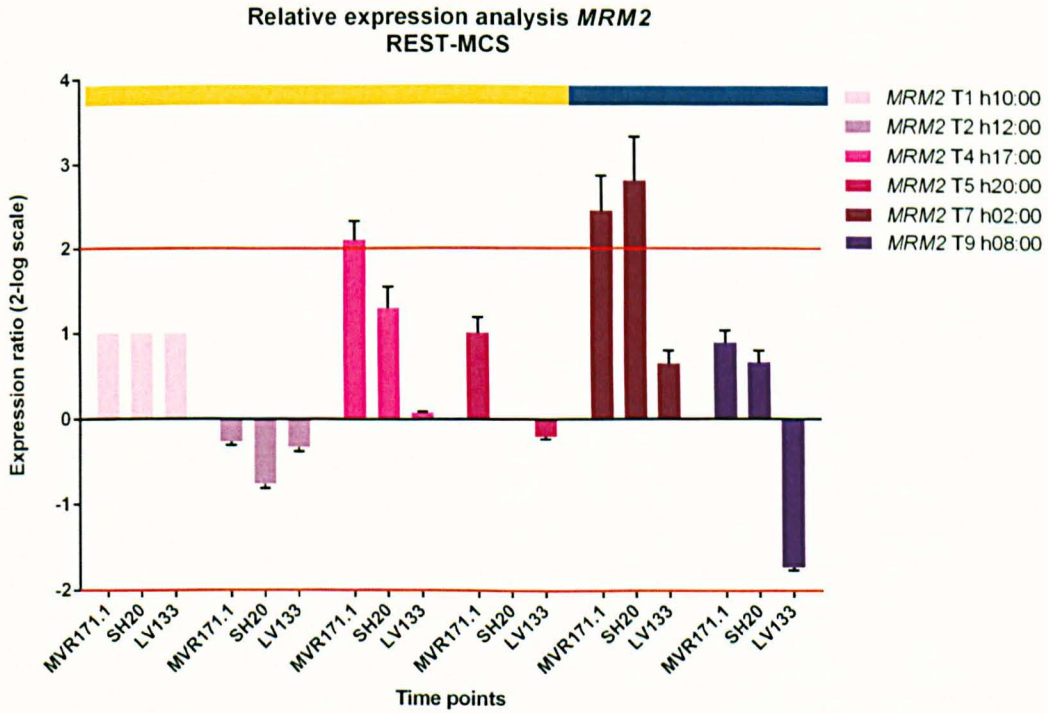


Figure D11: REST analysis of *MRM2* obtained by fixing T1 as reference condition; normalized against two reference genes *CDK* and *TBP*.

MRM1 did not show any expression variation along the 24 h course. *MRM2* displayed a significant up-regulation in T7, for only two samples, with respect to T1 set as control. *MRM2* presented, so far, a uniform expression trend along the 24 h with only one spot of up-regulation at 02:00 am.

Legatoria - Tipografia RICCO - Via Mezzocannone, 105 Napoli
Tel./Fax 081.551.66.55 e-mail: tipografiaricco@libero.it
Premiata con Medaglia d'oro "Premio Minerva 1988"